# Table of Content

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning</td>
<td>1</td>
</tr>
<tr>
<td>An Efficient and Generic Method for Interpreting Deep Learning based Knowledge Tracing Models</td>
<td>2</td>
</tr>
<tr>
<td>Deliang Wang, Yu Lu, Zhi Zhang and Penghe Chen</td>
<td></td>
</tr>
<tr>
<td>Recommending Learning Actions Using Neural Network</td>
<td>12</td>
</tr>
<tr>
<td>Hirokazu Kohama, Yuki Ban, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, Akitoshi Itai and Hiroyasu Usami</td>
<td></td>
</tr>
<tr>
<td>An Agent-Based Modeling and Simulation Tool as a Learning Aid for Diffusion of Innovations</td>
<td>21</td>
</tr>
<tr>
<td>Joseph Benjamin Ilagan, Jose Ramon Ilagan and Maria Mercedes Rodrigo</td>
<td></td>
</tr>
<tr>
<td>Studying Memory Decay and Spacing within Knowledge Tracing</td>
<td>24</td>
</tr>
<tr>
<td>Cristina Maier, Isha Slavin, Ryan S. Baker and Steve Stalzer</td>
<td></td>
</tr>
<tr>
<td>Developing a Video-based e-Learning System Incorporating a Fill-in-the-blank Question-type Concept Map</td>
<td>34</td>
</tr>
<tr>
<td>Takumi Hasegawa and Tessai Hayama</td>
<td></td>
</tr>
<tr>
<td>Exploring Students' Adoption of ChatGPT as a Mentor for Undergraduate Computing Projects: PLS-SEM Analysis</td>
<td>40</td>
</tr>
<tr>
<td>Swapna Gottipati, Kyong Jin Shim and Venky Shankararaman</td>
<td></td>
</tr>
<tr>
<td>Deep Knowledge Tracing is an Implicit Dynamic Multidimensional Item Response Theory Model</td>
<td>46</td>
</tr>
<tr>
<td>Jill-Jênn Vie and Hisashi Kashima</td>
<td></td>
</tr>
<tr>
<td>Immediate Feedback in Computational Thinking: Generating hints using a Knowledge Graph</td>
<td>55</td>
</tr>
<tr>
<td>Nitesh Kumar Jha, Plaban Kumar Bhowmik and Kaushal Kumar Bhagat</td>
<td></td>
</tr>
<tr>
<td>Composite Score for ChatGPT Prompt Efficiency: A Computational Linguistic Analysis of Engineered Chatbot Prompts</td>
<td>58</td>
</tr>
<tr>
<td>Michelle Banawan</td>
<td></td>
</tr>
<tr>
<td>Large Language Models (GPT) for Automating Feedback on Programming Assignments</td>
<td>68</td>
</tr>
<tr>
<td>Maciej Pankiewicz and Ryan Baker</td>
<td></td>
</tr>
<tr>
<td>Improved Automated Labeling of Mathematical Exercises in Japanese</td>
<td>78</td>
</tr>
<tr>
<td>Taisei Yamauchi, Ryosuke Nakamoto, Yiling Dai, Kyosuke Takami, Brendan Flanagan and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Can We Ensure Accuracy and Explainability for a Math Recommender System?</td>
<td>88</td>
</tr>
<tr>
<td>Yiling Dai, Brendan Flanagan and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Activity Analysis Support System by Causal Relation Check Between Sub-activities</td>
<td>98</td>
</tr>
<tr>
<td>Kota Kunori and Tomoko Kojiri</td>
<td></td>
</tr>
</tbody>
</table>
# Table of Content

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method for Estimating Learning Strategies from Tools Using Bayesian Network</td>
<td>101</td>
</tr>
<tr>
<td>Kento Kuwajima, Atsushi Ashida and Tomoko Kojiri</td>
<td></td>
</tr>
<tr>
<td>ExGen: Ready-To-Use Exercise Generation in Introductory Programming Courses</td>
<td>104</td>
</tr>
<tr>
<td>Nguyen Binh Duong Ta, Hua Gia Phuc Nguyen and Swapna Gottipati</td>
<td></td>
</tr>
<tr>
<td>Object Identification Training Support System for Object-Oriented Design with Cooking Recipes</td>
<td>114</td>
</tr>
<tr>
<td>Daiki Maeda, Kota Kunori and Tomoko Kojiri</td>
<td></td>
</tr>
<tr>
<td>Development of Learning Support System for Critical Reading of Academic Papers</td>
<td>120</td>
</tr>
<tr>
<td>Aota Nishida, Kazuhisa Seta and Yuki Hayashi</td>
<td></td>
</tr>
<tr>
<td>Overcoming Barriers to Sustainable Dissemination of L2 Learning Resources: An Integrated Framework for Creating and Distributing Dialogue Scenarios</td>
<td>123</td>
</tr>
<tr>
<td>Emmanuel Ayedoun, Yuki Hayashi and Kazuhisa Seta</td>
<td></td>
</tr>
<tr>
<td>Learning Support System to Understand Others Through Dramatic Script Reading and Its Evaluation</td>
<td>126</td>
</tr>
<tr>
<td>Hanano Okamoto, Yuki Hayashi and Kazuhisa Seta</td>
<td></td>
</tr>
<tr>
<td>Development of Estimation Method for Learner’s Emotional Concealment During Learning Using Biometric Information and Feedback Model</td>
<td>132</td>
</tr>
<tr>
<td>Koichi Shinohara, Keiichi Muramatsu and Tatsunori Matsui</td>
<td></td>
</tr>
<tr>
<td>Development of a Learning Support System for playing Ryuteki in Gagaku for Beginners</td>
<td>135</td>
</tr>
<tr>
<td>Yasushi Ueno and Masato Soga</td>
<td></td>
</tr>
<tr>
<td>An Adaptive Learning Support System based on Ontology of Multiple Programming Languages</td>
<td>141</td>
</tr>
<tr>
<td>Lalita Na Nongkhai, Jingyun Wang and Takahiko Mendori</td>
<td></td>
</tr>
<tr>
<td>Examination of the Robot’s Role as a Helper in Learning Situations</td>
<td>144</td>
</tr>
<tr>
<td>Makoto Shiraishi and Tatsunori Matsui</td>
<td></td>
</tr>
<tr>
<td>A Bio-Inspired Method for Personalized Learning Path Recommendation Problem</td>
<td>147</td>
</tr>
<tr>
<td>Yaqian Zheng, Deliang Wang, Yaping Xu, Ziqi Mao, Yaqi Zhao and Yanyan Li</td>
<td></td>
</tr>
<tr>
<td>Developing Effective Educational Chatbots with ChatGPT: Insights from Preliminary Tests in a Case Study on Social Media Literacy</td>
<td>150</td>
</tr>
<tr>
<td>Cansu Koyuturk, Mona Yavari, Emily Theophilou, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Alessandro Gabbiadini, Davinia Hernandez-Leo, Martin Ruskov and Dimitri Ognibene</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Tackling Unserious Raters in Peer Evaluation: Behavior Analysis and Early Detection with Learner Model</td>
<td>154</td>
</tr>
<tr>
<td>Changhao Liang, Izumi Horikoshi, Rwitajit Majumdar and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Future City: A Simulation for Making SDGs Action Decisions</td>
<td>164</td>
</tr>
<tr>
<td>Pin-Chen Chen, Yu-Hao Lu and Ju-Ling Shih</td>
<td></td>
</tr>
<tr>
<td>Multimodal Assessment of an Ultra-Brief Practice of Progressive Muscular Relaxation Adapted for the Classroom</td>
<td>167</td>
</tr>
<tr>
<td>Marc Beardsley, Batuhan Sayis and Marta Portero Tresserra</td>
<td></td>
</tr>
<tr>
<td>The Development and Evaluation of the Platform for Online High-Level Cooperative Games</td>
<td>177</td>
</tr>
<tr>
<td>Feng-Lung Liu, Geng-De Hong, Ju-Ling Shih and George Ghinea</td>
<td></td>
</tr>
<tr>
<td>A Thematic Analysis Exploring Flexibility in Programming-based Mathematical Problem Solving</td>
<td>183</td>
</tr>
<tr>
<td>Huiyan Ye, Oi-Lam Ng and Zhihao Cui</td>
<td></td>
</tr>
<tr>
<td>Epistemic Network Analysis to assess collaborative engagement in Knowledge Building discourse</td>
<td>189</td>
</tr>
<tr>
<td>Aloysius Ong, Chew Lee Teo, Alwyn Vwen Yen Lee and Guangji Yuan</td>
<td></td>
</tr>
<tr>
<td>Investigating Trustworthiness and Conflict in Historical Multiple Texts: From Eye-Tracking Data of Source and Content Processing</td>
<td>199</td>
</tr>
<tr>
<td>Zheng-Hong Guan and Sunny S. J. Lin</td>
<td></td>
</tr>
<tr>
<td>Multimodal Analysis of Learners' Communications in CSCL of a Mathematical Proof</td>
<td>206</td>
</tr>
<tr>
<td>Masataka Kaneko, Hironori Egi and Takeo Noda</td>
<td></td>
</tr>
<tr>
<td>Development and Evaluation of a Matching System to Facilitate Online Collaborative Learning</td>
<td>209</td>
</tr>
<tr>
<td>Haruka Tada and Fumihide Tanaka</td>
<td></td>
</tr>
<tr>
<td>Assessing college students’ sense of community for advancing community knowledge</td>
<td>219</td>
</tr>
<tr>
<td>Chih Hui Seet, Yi-Ning Tsai, Guo-Tsai Hung and Huang-Yao Hong</td>
<td></td>
</tr>
<tr>
<td>Teachers Developing More Creative Learning Views via Online Perspective Taking Activities</td>
<td>222</td>
</tr>
<tr>
<td>Mei-Ju Chen, Chao-Yu Guo and Huang-Yao Hong</td>
<td></td>
</tr>
<tr>
<td>Unveiling University Students' Data Literacy: A Case Study on Modeling Reasoning in Data Mining Projects</td>
<td>225</td>
</tr>
<tr>
<td>Tianqi Zhang</td>
<td></td>
</tr>
<tr>
<td>Mathematic Learning-by-teaching: Video Creation and Cross-Schools Staging</td>
<td>235</td>
</tr>
<tr>
<td>Te-Yang Chou, Yen-Cheng Yeh and Hong-Ming TU</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>The Role of Individual Preparation for Knowledge Construction in Collaborative Argumentation: an Epistemic Network Analysis</td>
<td>238</td>
</tr>
<tr>
<td><em>Wenli Chen, Junzhu Su, Qianru Lyu, Siew Cheng Aileen Chai, Xinyi Li, Guo Su and Eng Eng Ng</em></td>
<td></td>
</tr>
<tr>
<td>Students Know more than They can Tell: Understanding Learners’ Ideas of Heat Transfer via Model Revision Activities</td>
<td>244</td>
</tr>
<tr>
<td><em>Rajashri Priyadarshini, Chandan Dasgupta and Sahana Murthy</em></td>
<td></td>
</tr>
<tr>
<td>Impact of 360°VR on Pre-Service Teachers’ Empathy——Taking Educational Equity as an example</td>
<td>250</td>
</tr>
<tr>
<td><em>Yanjun Chen, Yiling Hu and Bian Wu</em></td>
<td></td>
</tr>
<tr>
<td>Enhancing Student Teachers’ Collaborative Interdisciplinary Design Through Knowledge-Building Activities</td>
<td>253</td>
</tr>
<tr>
<td><em>Pei-Yi Lin</em></td>
<td></td>
</tr>
<tr>
<td>From Individual Ideation to Group Knowledge Co-Construction: Comparison of High- and Low-performing Groups</td>
<td>256</td>
</tr>
<tr>
<td><em>Wenli Chen, Guo Su, Xinyi Li, Qianru Lyu, Junzhu Su, Siew Cheng Aileen Chai and Eng Eng Ng</em></td>
<td></td>
</tr>
<tr>
<td>Discussion Support Agent System to Promote Equalization of Speech among Participants</td>
<td>266</td>
</tr>
<tr>
<td><em>Ryunosuke Nishimura, Risa Iharada, Yuya Sugamoto, Yutaka Ishii, Toshio Mochizuki and Hironori Egi</em></td>
<td></td>
</tr>
<tr>
<td>Experimental Verification of &quot;Peer-ness&quot; Formation by a Learning Companion Robot — Possibility of inducing a sense of competition through long-term nonverbal interaction—</td>
<td>272</td>
</tr>
<tr>
<td><em>Koki Honda, Yoshimasa Tawatsui and Tatsunori Matsui</em></td>
<td></td>
</tr>
<tr>
<td>Study on The Development of Computational Thinking Decomposition Strategies for Senior Primary Students</td>
<td>278</td>
</tr>
<tr>
<td><em>Mengtao Li, Yaxin Guan and Guang Chen</em></td>
<td></td>
</tr>
<tr>
<td>Remembering The Knowledge of Experts and Novices in Computer-Supported Collaborative Learning: A Multinomial Processing Tree Approach</td>
<td>286</td>
</tr>
<tr>
<td><em>Oktay Ülker and Daniel Bodemer</em></td>
<td></td>
</tr>
<tr>
<td>Exploring Group Formation Strategies in Computer-Supported Collaborative Learning: A Systematic Review</td>
<td>292</td>
</tr>
<tr>
<td><em>Jiamin Tang, Huihan Zhou, Yajing Tan and Guang Chen</em></td>
<td></td>
</tr>
<tr>
<td>A Comparative Analysis on the Effects of Cognitive Tools in Data Inquiry Cultivation</td>
<td>295</td>
</tr>
<tr>
<td><em>Hui Zhang, Bian Wu, Yiling Hu and Yujie Xu</em></td>
<td></td>
</tr>
<tr>
<td>Argumentative Knowledge Construction and Certainty Navigation: A Comparison between Individual and Group Work</td>
<td>301</td>
</tr>
<tr>
<td><em>Wenli Chen, Eng Eng Ng, Guo Su, Junzhu Su, Xinyi Li, Siew Cheng Aileen Chai and Qianru Lyu</em></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>ChatGPT’s Performance in Spreadsheets Modeling Assessments based on Revised Bloom’s Taxonomy</td>
<td>307</td>
</tr>
<tr>
<td>Michelle Cheong</td>
<td></td>
</tr>
<tr>
<td>Promoting Middle School Students' Achievement and Attitude toward Science Learning through Sphere Recognition-Based AR Application</td>
<td>308</td>
</tr>
<tr>
<td>Ruixue Liu, Lijun Liang and Xiaodong Wei</td>
<td></td>
</tr>
<tr>
<td>Preparation for Future Lockdowns: A Comparison of Student LMS Activity During and After COVID-19</td>
<td>318</td>
</tr>
<tr>
<td>Ezekiel Adriel Lagmay and Maria Mercedes T. Rodrigo</td>
<td></td>
</tr>
<tr>
<td>Using Augmented Reality to Facilitate Music Learning for Preschool Children</td>
<td>328</td>
</tr>
<tr>
<td>Xiaodong Wei, Rui Qiu and Ruixue Liu</td>
<td></td>
</tr>
<tr>
<td>Comparing Perceived Cognitive Load while Learning Online with AI Chatbots, Pre-recorded Videos, and Live Lectures</td>
<td>338</td>
</tr>
<tr>
<td>Haixi Sheng, Xinran Zhou, Yue Zhao and Guoqing Zhao</td>
<td></td>
</tr>
<tr>
<td>Do the Same Rules Apply? Transferring MOOC Success Behaviors to University Online Learning</td>
<td>344</td>
</tr>
<tr>
<td>Clarence James Monterozo and Maria Mercedes Rodrigo</td>
<td></td>
</tr>
<tr>
<td>Visual Attention Patterns in Processing Compiler Error Messages</td>
<td>350</td>
</tr>
<tr>
<td>Christine Lourrine Tablatin and Maria Mercedes Rodrigo</td>
<td></td>
</tr>
<tr>
<td>Process Evaluation for Concept Map Building and Its Experimental Evaluation</td>
<td>356</td>
</tr>
<tr>
<td>Ridwan Rismanto, Aryo Pinandito, Banni Satria Andoko, Yusuke Hayashi and Tsukasa Hirashima</td>
<td></td>
</tr>
<tr>
<td>On-demand lectures with humor and questions using avatars</td>
<td>362</td>
</tr>
<tr>
<td>Satoshi Toyota and Asuka Terai</td>
<td></td>
</tr>
<tr>
<td>Adapting Noticing Framework to Analyze Learner's Reasoning in VR-simulated complex scenarios</td>
<td>372</td>
</tr>
<tr>
<td>Herold Pc and Chandan Dasgupta</td>
<td></td>
</tr>
<tr>
<td>Analysis of Algorithmic Strategy Development in the Development of Computational Thinking of Upper Elementary School Students</td>
<td>382</td>
</tr>
<tr>
<td>Xiaowen Wang, Pinqi Hu and Guang Chen</td>
<td></td>
</tr>
<tr>
<td>Toward a Virtual Human Exhibit for Public AI Education</td>
<td>388</td>
</tr>
<tr>
<td>Ning Wang, Timothy Hurt, Ari Krakowski, Eric Greenwald, Omkar Masur, Boxi Fu and Chirag Merchant</td>
<td></td>
</tr>
</tbody>
</table>
## Table of Content

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzing Learning Patterns and Potential Interventions in First-Year Compulsory Course at an Online University</td>
<td>397</td>
</tr>
<tr>
<td>Yasuhisa Kato</td>
<td></td>
</tr>
<tr>
<td>A Comparative Study of Traditional and Augmented Reality-Based Engineering Drawing Instruction: Effects on Visualization Skills and Cognitive Load</td>
<td>400</td>
</tr>
<tr>
<td>Ajay Shankar Tiwari and Kaushal Kumar Bhagat</td>
<td></td>
</tr>
<tr>
<td>Measuring Self-regulated Learning Processes in Computer Science Education</td>
<td>406</td>
</tr>
<tr>
<td>Elizabeth Cloude, Ryan Baker and Maciej Pankiewicz</td>
<td></td>
</tr>
<tr>
<td>Construction of a Japanese Language Learning Support System for Learning Semantic Negotiation</td>
<td>409</td>
</tr>
<tr>
<td>Satoru Kogure, Akira Yoshida, Yasuhiro Noguchi, Koichi Yamashita, Tatsuhiro Konishi and Makoto Kondo</td>
<td></td>
</tr>
<tr>
<td>Creating Meaningful Connections: The Role of Simultaneous Multi Situational Learning in Knowledge Contextualization and Application</td>
<td>412</td>
</tr>
<tr>
<td>Chih-Yen Chen, Su-Hang Yang, Meng-Xuan Xie, Yi-Chuan Fan, Jen-Hang Wang and Gwo-Dong Chen</td>
<td></td>
</tr>
<tr>
<td>ECLAIR: A Centralized AI-Powered Recommendations System in a Multi-Node EXAIT System</td>
<td>422</td>
</tr>
<tr>
<td>Isanka Wijerathne, Brendan Flanagan, Yiling Dai and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Sharing Learning Log while Maintaining Privacy over Blockchain: Heuristic Evaluation of BOLL</td>
<td>429</td>
</tr>
<tr>
<td>Patrick Ocheja, Rwitajit Majumdar, Brendan Flanagan and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Program Visualization System Supporting Teacher-Intended Stepwise Refinement</td>
<td>435</td>
</tr>
<tr>
<td>Koichi Yamashita, Hiroki Soma, Satoru Kogure, Yasuhiro Noguchi, Raiya Yamamoto, Tatsuhiro Konishi and Yukihiro Itoh</td>
<td></td>
</tr>
<tr>
<td>Automatic Detection of Negotiation in Collaborative Complex Problem Solving Interactions</td>
<td>441</td>
</tr>
<tr>
<td>Daevesh Singh, Ulfja Khwaja, Sahana Murthy and Ramkumar Rajendran</td>
<td></td>
</tr>
<tr>
<td>A page jump recommendation model based on digital textbook contents and student log data</td>
<td>451</td>
</tr>
<tr>
<td>Wenhao Wang, Natsumi Yamamoto, Fuzheng Zhao, Etsuko Kumamoto, Zicheng Kang and Chengjiu Yin</td>
<td></td>
</tr>
<tr>
<td>Concept Map Recomposition Approach for Advanced Formative Assessment in Large-Scale Online Course</td>
<td>457</td>
</tr>
<tr>
<td>Sho Yamamoto, Aryo Pinandito and Tsukasa Hirashima</td>
<td></td>
</tr>
<tr>
<td>Towards Automated Evidence Extraction: A Case of Adapting SAM to Real-World Educational Data</td>
<td>463</td>
</tr>
<tr>
<td>Koki Okumura, Izumi Horikoshi, Kento Koike and Hiroaki Ogata</td>
<td></td>
</tr>
</tbody>
</table>
## Table of Content

What does process mining of feedback-seeking reveal about problem-solving in chemistry undergraduates?  
*Anveshna Srivastava and Chandan Dasgupta*  
472

ICCE 2023 Learning Outcomes of Computer Programming and Information Technology - Integrated Courses for Non-Computer Science Majors: Case Study of a Public Research University in Taiwan  
*Che-Yu Hsu, Feng-Nan Hwang, Tseng-Yi Chen and Chia-Hui Chang*  
475

DLOT: An Open-Source Application to Assist Human Observers  
*Ashwin T S, Danish Shafi Shaikh and Ramkumar Rajendran*  
478

Supporting Learning Through Affordance-Based Design: A Comparative Analysis of "BioVARse" and a Standard Textbook Companion Application in Biology Education  
*Devanshu Saindane, Sunny Prakash Prajapati and Syaamantak Das*  
484

Investigating Programming Performance Predictability from Embedding Vectors of Coding Behaviors  
*Ikkei Igawa, Yuta Taniguchi, Tsubasa Minematsu, Fumiya Okubo and Atsushi Shimada*  
487

Unveiling Learners’ Interaction Behavior in Virtual Reality Learning Environment  
*Antony Prakash and Ramkumar Rajendran*  
490

Keeping Teams in the Game: Predicting Dropouts in Online Problem-Based Learning Competition  
*Aditya Panwar, Ashwin T S, Ramkumar Rajendran and Kavi Arya*  
500

### C4: ICCE Sub-Conference on Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)

Fostering Students’ Dialogic Engagement with the Use of Visual Learning Analytics as a Teaching Assistant Tool in Primary School Classrooms  
*Fan Chen, Pengjin Wang, Deliang Wang and Gaowei Chen*  
511

Advancing Education through Stakeholder Engagement: An Evaluation of the Learning Butler Chatbot's Impact on Instructors, and Learners  
*Chih-Yang Peng, Su-Hang Yang, Pei-Yu Ho, Jen-Hang Wang and Gwo-Dong Chen*  
521

Building Students’ Learning Habits on Slack: An Application of the IDC Theory  
*Veenita Shah, Sahana Murthy and Sridhar Iyer*  
527

Using the Self-regulated Based Personalized Online Learning System for Learning Factorization in Mathematics  
*Thanyaluck Ingkavara, Patcharin Panjaburee and Wararat Wongkia*  
537
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Effects of Visualization Strategies on Students’ Learning Outcomes in Augmented Reality Contexts</td>
<td>543</td>
</tr>
<tr>
<td>Pei-Shan Tsai</td>
<td></td>
</tr>
<tr>
<td>Teaching Analytics with xAPI: Learning Activity Visualization with Cross-platform Data</td>
<td>548</td>
</tr>
<tr>
<td>Izumi Horikoshi, Yuko Toyokawa, Kohei Nakamura, Changhao Liang, Rwitajit Majumdar and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Effects of Audio and Tactile Biofeedback Based on EEG Attention Levels on University Students’ Relaxation</td>
<td>554</td>
</tr>
<tr>
<td>Shih-Jou Yu, Wai Ki Rebecca Cheng, Yi-Hsuan Chen and Jerry Chih-Yuan Sun</td>
<td></td>
</tr>
<tr>
<td>Estimating Physical Interactions with Neighboring Student for Detecting Active Learners in the Computer Classroom</td>
<td>560</td>
</tr>
<tr>
<td>Takahiro Yoshino, Shin Ueno and Hironori Egi</td>
<td></td>
</tr>
<tr>
<td>Effect of Active Breaks during e-Learning and Mental Arithmetic Tasks</td>
<td>563</td>
</tr>
<tr>
<td>Masaki Kodaira, Tatsuya Hamada and Hironori Egi</td>
<td></td>
</tr>
<tr>
<td>Chronotypes of Learning Habits in Weekly Math Learning of Junior High School</td>
<td>566</td>
</tr>
<tr>
<td>Chia-Yu Hsu, Mandukhai Otgonbaatar, Izumi Horikoshi, Huiyong Li, Rwitajit Majumdar and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>C5: ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)</td>
<td>569</td>
</tr>
<tr>
<td>Relationship Between Students’ Minecraft Re-engagement Metrics and STEM Interest</td>
<td>570</td>
</tr>
<tr>
<td>Maria Mercedes T. Rodrigo, Jonathan Casano and Mikhail Fuentes</td>
<td></td>
</tr>
<tr>
<td>Enhancing Learning Experience in University Engineering Classes with Kahoot! Quiz Games</td>
<td>573</td>
</tr>
<tr>
<td>Zilu Liang</td>
<td></td>
</tr>
<tr>
<td>Design and Implementation of an Educational Game for Teaching Artificial Intelligence to High School Students</td>
<td>579</td>
</tr>
<tr>
<td>Ning Wang, Ryan Montgomery, Eric Greenwald and Maxyn Leitner</td>
<td></td>
</tr>
<tr>
<td>Executive Functions Training-oriented Digital Games: Effectiveness and Experience</td>
<td>589</td>
</tr>
<tr>
<td>Chi-Fang Huang, Hui-Ling Hsu and Zhi-Hong Chen</td>
<td></td>
</tr>
<tr>
<td>Incorporating Tangible Rewards into Gamification Increases Students’ Identified Regulation in Fully Online Learning</td>
<td>595</td>
</tr>
<tr>
<td>Ya Xiao and Khe Foon Hew</td>
<td></td>
</tr>
<tr>
<td>Effects of Different Embodied Scaffoldings on Students' Spatial Abilities in Digital Game-Based Learning</td>
<td>601</td>
</tr>
<tr>
<td>Junyi Zhou, Jialing Zeng, John B. Black and Junjie Shang</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Exploring the Impact of Designing a Robot as a Pet with Interdependence Theory on Long-Term Relationships and Learning Performance</td>
<td>611</td>
</tr>
<tr>
<td>Vando Gusti Al Hakim, Su-Hang Yang, Jen-Hang Wang, Yu-Chen Chang, Hung-Hsuan Lin and Gwo-Dong Chen</td>
<td></td>
</tr>
<tr>
<td>Analysis to Creation: Using the ADDIE Model to Develop an Educational Game for Children</td>
<td>621</td>
</tr>
<tr>
<td>Yufan Zhang, Nurul Nadwa Zulkifli, Ahmad Fauzi Mohd Ayub and Zewen Shang</td>
<td></td>
</tr>
<tr>
<td>Case Study-Based Research on Understanding App User Engagement to Develop Environmental Literacy of Urban Residents</td>
<td>624</td>
</tr>
<tr>
<td>Ewa Duda</td>
<td></td>
</tr>
<tr>
<td>Design and Development of A Game to Improve Self-Efficacy: A Case Study of Addressing Modes Learning</td>
<td>627</td>
</tr>
<tr>
<td>Fuzheng Zhao, Danqing Luo, Etsuko Kumamoto and Chengjiu Yin</td>
<td></td>
</tr>
<tr>
<td>Investigating Cognitive Biases in Self-Explanation Behaviors during Game-based Learning about Mathematics</td>
<td>637</td>
</tr>
<tr>
<td>J. M. Alexandra Andres, Elizabeth Cloude, Ryan Baker and Seiyon Lee</td>
<td></td>
</tr>
<tr>
<td>GaMINLab - Meaningful Gamification to Engage Students in Science Inquiry Practices through Simulation Labs</td>
<td>643</td>
</tr>
<tr>
<td>Archana Rane, Sahana Murthy and Sasikumar M.</td>
<td></td>
</tr>
<tr>
<td>A Skill Tracing Model for Player Character Control in STG</td>
<td>646</td>
</tr>
<tr>
<td>Peizhe Huang, Wanxiang Li, Wen Gu, Kouichi Ota and Shinobu Hasegawa</td>
<td></td>
</tr>
<tr>
<td>The Design and Practice of Scientific Inquiry Activities for Children Aged 5-6 Based on an AR Flashcard Environment</td>
<td>649</td>
</tr>
<tr>
<td>Tian-Tian Gou, Min-Sheng Fan and Bin-Li Wang</td>
<td></td>
</tr>
<tr>
<td>Toward Game-Based Learning of Japanese Writing for Elementary School Students</td>
<td>655</td>
</tr>
<tr>
<td>Kazumasa Omura, Kei Kubo, Frederic Bergeron and Sadao Kurohashi</td>
<td></td>
</tr>
<tr>
<td>The Impact of Digital Game-based Learning with a Mathematical Game Application on Calculation Abilities of Grade 4 Students</td>
<td>661</td>
</tr>
<tr>
<td>Yinbei Liu and Alex Wing Cheung Tse</td>
<td></td>
</tr>
<tr>
<td>Develop and validate STEM education activities using the “6E Design Teaching Model”: Taking “Dynamics and Energy Conversion in Sail Car Design” as an example</td>
<td>671</td>
</tr>
<tr>
<td>Lu Chen, Yang Yang and Guang Chen</td>
<td></td>
</tr>
<tr>
<td>The Impact of Gamified Assessment on the Learning Burnout of Undergraduate Computing Students: a Quasi-experimental Research</td>
<td>686</td>
</tr>
<tr>
<td>Beilei Zhang and Alex Wing Cheung Tse</td>
<td></td>
</tr>
</tbody>
</table>
# Table of Content

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhancing Learner Satisfaction in Simulation-Based Learning: The Impact of Learner Characteristics and Expectancy</td>
<td>696</td>
</tr>
<tr>
<td>Steven Ck Ng, Li Fern Tan and Poh Nguk Lau</td>
<td></td>
</tr>
<tr>
<td>Exploring the Impact on Student Reading Preferences in Gamified Reading Portfolio</td>
<td>704</td>
</tr>
<tr>
<td>Hsiao-Tung Yang, Chang-Yen Liao, Ciao-Min Syu and Tak-Wai Chan</td>
<td></td>
</tr>
<tr>
<td>Improving Teamwork through a Decision-Theoretic Coach in a Minecraft Search-and-Rescue Game</td>
<td>707</td>
</tr>
<tr>
<td>David Pynadath, Nik Gurney, Sarah Kenny, Rajay Kumar, Stacy Marsella, Haley Matuszak, Hala Mostafa, Pedro Sequeira, Volkan Ustun and Peggy Wu</td>
<td></td>
</tr>
<tr>
<td>C6: ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)</td>
<td>710</td>
</tr>
<tr>
<td>Using Flipped Classroom Approach to Integrate SRL Instruction in Classical Chinese Reading Instruction: Insights from the First-Year Results</td>
<td>711</td>
</tr>
<tr>
<td>Kit Ling Lau</td>
<td></td>
</tr>
<tr>
<td>Relations between Instructional Factors and Student Acceptance of Flipped Learning in Chinese Language Learning</td>
<td>717</td>
</tr>
<tr>
<td>Kit Ling Lau and Quan Qian</td>
<td></td>
</tr>
<tr>
<td>The AI-Supported Instructional Design in PBL Integrating Chinese Language Learning and Multimedia Creation</td>
<td>726</td>
</tr>
<tr>
<td>Satoko Sugie</td>
<td></td>
</tr>
<tr>
<td>Experimental Comparison of Promotion Effect for EFL Reading Comprehension between Conventional Summarization and Toulmin Argument Reconstruction</td>
<td>732</td>
</tr>
<tr>
<td>Banni Satria Andoko, Rosa Andrie Asmara, Vivin Ayu Lestari, Deasy Sandhya Elya Ikawati, Arief Prasetyo, Tsukasa Hirashima and Yusuke Hayashi</td>
<td></td>
</tr>
<tr>
<td>Effects of a Machine Learning-empowered Chinese Character Handwriting Learning Tool on Rectifying Legible Writing in Young Children: A Pilot Study</td>
<td>742</td>
</tr>
<tr>
<td>Lung-Hsiang Wong, Guat Poh Aw, He Sun, Ching-Chiuan Yen, Chor Guan Teo and Yun Wen</td>
<td></td>
</tr>
<tr>
<td>Using TAASSC to Investigate Fine-Grained Complexity in Reading Texts of Two High-Stakes English Tests in China</td>
<td>748</td>
</tr>
<tr>
<td>Shengshu Lin</td>
<td></td>
</tr>
<tr>
<td>Using Multidimensional Analysis to Investigate the Extrapolation Inference of a High-Stakes Test</td>
<td>754</td>
</tr>
<tr>
<td>Shengshu Lin</td>
<td></td>
</tr>
<tr>
<td>Enabling Visually Impaired People’s Chinese Literacy learning through Information Technology</td>
<td>760</td>
</tr>
<tr>
<td>Shelley Shwu-Ching Young and Jen-Li Wang</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Tracking Knowledge for Learning Japanese as a 2nd Language</td>
<td>766</td>
</tr>
<tr>
<td>Tomoko Okimoto, Matthew Johnson, Huy Nguyen, Steven Moore, Michael Eagle and John Stamper</td>
<td></td>
</tr>
<tr>
<td>The Effect of Timing Differences in Online Corrective Feedback on Adult Verbal English Learners’ Learning Engagement: A Micro-genetic Study</td>
<td>769</td>
</tr>
<tr>
<td>Wanying Liang, Guang Chen and Wei Cheng</td>
<td></td>
</tr>
<tr>
<td>Using Self-Regulated Digital Storytelling in Primary Students’ English Learning: An Exploratory Factor Analysis</td>
<td>779</td>
</tr>
<tr>
<td>Yunsi Tina Ma, Siu Cheung Kong and Daner Sun</td>
<td></td>
</tr>
<tr>
<td>Constructing a Natural Conversation Learning Activity to Improve Students’ English-Speaking Skills</td>
<td>789</td>
</tr>
<tr>
<td>Hong-Min Tu, Chang-Yen Liao and Te-Yang Chou</td>
<td></td>
</tr>
<tr>
<td>Analysis of Topic Sentences of Classification Paragraphs and Development of a Diagnostic Function</td>
<td>792</td>
</tr>
<tr>
<td>Kandai Ishikawa and Hidenobu Kunichika</td>
<td></td>
</tr>
<tr>
<td>Construction of an English Grammar Quiz Recommendation System Using Explanation by a Knowledge Map</td>
<td>795</td>
</tr>
<tr>
<td>Kensuke Takii, Naomichi Tanimura, Brendan Flanagan and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Educational System of English Tense for Japanese Learners by Forming Temporal Constraints on Tense</td>
<td>801</td>
</tr>
<tr>
<td>Haruto Nagata, Tatsuhiro Konishi and Makoto Kondo</td>
<td></td>
</tr>
<tr>
<td>Exploring the Use of Chatbot to Promote Online EFL Students’ Behavioral, Cognitive, and Emotional Engagements</td>
<td>807</td>
</tr>
<tr>
<td>Xinyi Luo, Weijiao Huang, Khe Foon Hew, Chengyuan Jia and Xiangjie Cao</td>
<td></td>
</tr>
<tr>
<td>Impact of Self-analysis Behaviors in GOAL for Japanese High School EFL Learners</td>
<td>813</td>
</tr>
<tr>
<td>Zixu Wang, Rwitajit Majumdar and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>AI-powered Collaborative Activities for Chinese Vocabulary Learning</td>
<td>819</td>
</tr>
<tr>
<td>Xinyu Guo and Yun Wen</td>
<td></td>
</tr>
<tr>
<td>Design and Development of a Sentence Construction Game for Deaf and Hard of Hearing (DHH) Users: A Qualitative Usability Study</td>
<td>825</td>
</tr>
<tr>
<td>Arjun Prasad, Sunny Prajapati, Utkarsh and Vishwas Badhe</td>
<td></td>
</tr>
<tr>
<td>Using T-Robot Board Games to Enhance Learning Gains of Rural Elementary School Students</td>
<td>831</td>
</tr>
<tr>
<td>Yu-Wei Wu, Wen-Yu Ye, Wen-Chi Vivian Wu and Rong-Jyue Wang</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>C7: ICCE Sub-Conference on Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)</td>
<td>834</td>
</tr>
<tr>
<td>How Teachers’ Conceptions of Student Engagement Influenced their Actual Strategy and Student Online Engagement</td>
<td>835</td>
</tr>
<tr>
<td>Gulipari Maimaiti and Khe Foon Hew</td>
<td></td>
</tr>
<tr>
<td>ICCE 2023 Exploring the Social Media Discourse: the Impact of ChatGPT on Teachers' Roles and Identity</td>
<td>838</td>
</tr>
<tr>
<td>Yuchun Zhong, Davy Tsz Kit Ng and Samuel Kai Wah Chu</td>
<td></td>
</tr>
<tr>
<td>Designing Faculty Development Programs by a Team from Different Majors</td>
<td>849</td>
</tr>
<tr>
<td>Yukari Kato, Yukari Yukihiko, Hisashi Hatakeyama and Toshiya Oishi</td>
<td></td>
</tr>
<tr>
<td>What Digital Tools Teachers are Ready to Use in Kindergarten – International Comparative Study with Early Childhood Pre-Service Teachers</td>
<td>855</td>
</tr>
<tr>
<td>Democratising AI Education: Teaching Autoencoders to Out-of-School Children from Low-income Backgrounds</td>
<td>861</td>
</tr>
<tr>
<td>Saumay Garg</td>
<td></td>
</tr>
<tr>
<td>Informatics Education for University Students based on Text Input Time</td>
<td>867</td>
</tr>
<tr>
<td>Yuko Murakami and Tomohiro Inagaki</td>
<td></td>
</tr>
<tr>
<td>Investigating Student Teachers’ Learning Experience and Choice of Participation Modalities in a HyFlex Course: A Mixed Methods Approach</td>
<td>873</td>
</tr>
<tr>
<td>Liang Jing Teh, Su Luan Wong, Mohd Zariat Abdul Rani, Mas Nida MD Khambari and Sai Hong Tang</td>
<td></td>
</tr>
<tr>
<td>Assessment of Intelligent Teaching Preparation of EFL Teachers: Based on Two-Year Data Comparison</td>
<td>884</td>
</tr>
<tr>
<td>Xin An, Xi Shen, Jiannan Bai and Yushun Li</td>
<td></td>
</tr>
<tr>
<td>Using Learning Design Technologies for Teachers’ Practice-Driven Research</td>
<td>892</td>
</tr>
<tr>
<td>Marc Beardsley, Davinia Hernandez-Leo and Roberto Sánchez-Reina</td>
<td></td>
</tr>
<tr>
<td>The Impact of Preschool Teacher Training on STEAM Education: Professional Preparedness and Confidence</td>
<td>898</td>
</tr>
<tr>
<td>Wan-Chen Chang</td>
<td></td>
</tr>
<tr>
<td>Matching Intervention Messages Considering Complex Personality Types of High School Students</td>
<td>901</td>
</tr>
<tr>
<td>Taisei Yamauchi, Yuta Nakamizo, Kyosuke Takami, Rwitajit Majumdar and Hiroaki Ogata</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Design-Based Implementation Research: A Collaborative Approach to Educate Out-of-School Children</td>
<td>907</td>
</tr>
<tr>
<td><em>Faisal Badar and Jon Mason</em></td>
<td></td>
</tr>
<tr>
<td>A Study of Versatile Tutor Training Programs for Universities in Japan</td>
<td>910</td>
</tr>
<tr>
<td><em>Yumi Ishige, Kazuhiro Kabeya, Kayoko Nagao and Hirotoshi Tanigawa</em></td>
<td></td>
</tr>
<tr>
<td>Support for Fitting Chromebooks to the Child with Cerebral Palsy: A Practical Study on Incorporating Advice from ICT Specialists</td>
<td>913</td>
</tr>
<tr>
<td><em>Tomohito Yamazaki and Toyokazu Mizuuchi</em></td>
<td></td>
</tr>
<tr>
<td>Exploring the Relationship between 21st Century Skills and Motivation: A Study Using Contextual Inquiry Project-based Learning</td>
<td>916</td>
</tr>
<tr>
<td><em>Jirapipat Thanyaphongphat, Preeyada Tapingkae, Kannika Daungcharone and Krittawaya Thongkoo</em></td>
<td></td>
</tr>
<tr>
<td>Data-Driven Competency Assessment Supporting System for Teachers</td>
<td>926</td>
</tr>
<tr>
<td><em>Taito Kano, Izumi Horikoshi, Kento Koike and Hiroaki Ogata</em></td>
<td></td>
</tr>
<tr>
<td>Visualization of Instructional Patterns from Daily Teaching Log Data</td>
<td>936</td>
</tr>
<tr>
<td><em>Kohei Nakamura, Izumi Horikoshi, Rwitajit Majumdar and Hiroaki Ogata</em></td>
<td></td>
</tr>
<tr>
<td>Sustaining Students’ Interest in an Instructional System Design Course by Leveraging Interest-Driven Creator Theory</td>
<td>946</td>
</tr>
<tr>
<td><em>Vishwas Badhe, Sunita Raste, Sahana Murthy and Sridhar Iyer</em></td>
<td></td>
</tr>
<tr>
<td>Supporting Science Teachers to Select Quality Edtech Learning Solutions for their Context</td>
<td>956</td>
</tr>
<tr>
<td><em>Shruti Jain, Sheeja Vasudevan, Madhuri Mavinkurve and Sahana Murthy</em></td>
<td></td>
</tr>
<tr>
<td>A Preliminary Study: Exploring Teachers’ Perspectives on the Role of Gathering Information in Supporting Teachers’ Digital Learning Agility</td>
<td>966</td>
</tr>
<tr>
<td><em>Kamilah Abdullah, Mas Nida Md Khambari, Nur Dania Mohd Rosli, Su Luan Wong, Noor Syamilah Zakaria, Siti Raba’Ah Hamzah and Priscilla Moses</em></td>
<td></td>
</tr>
<tr>
<td>The Role of Flexibility in Shaping Teachers’ Digital Learning Agility: A Preliminary Study</td>
<td>972</td>
</tr>
<tr>
<td><em>Nur Dania Mohd Rosli, Kamilah Abdullah, Mas Nida Md. Khambari, Su Luan Wong, Noor Syamilah Zakaria, Priscilla Moses and Siti Raba’Ah Hamzah</em></td>
<td></td>
</tr>
<tr>
<td>Incorporation of Robotics in AI Education: Barriers and Enablers as Perceived by K-12 Teachers</td>
<td>978</td>
</tr>
<tr>
<td><em>Miao Yue, Siu-Yung Jong, Yun Dai, Tsun-Hin Luk and Ma. Jenina N. Nalipay</em></td>
<td></td>
</tr>
<tr>
<td>Collaborative Design of a Simulation-based Math Classroom: Contradictions and Solutions between Teaching and Research</td>
<td>981</td>
</tr>
<tr>
<td><em>Wenxin Guo, Bian Wu and Dong Li</em></td>
<td></td>
</tr>
<tr>
<td>The Impact of Development-questioning Activities on Students' Pre-writing Ideas</td>
<td>991</td>
</tr>
<tr>
<td><em>Jia Ling Hong, Chang-Yen Liao, Tak-Wai Chan and Jui-Fen Chang</em></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Promoting Teachers’ Digital Literacy Achievement: A Nationwide Survey of Education</td>
<td>997</td>
</tr>
<tr>
<td>Ziyan Che, Jiumei Yang, Longkai Wu and Di Wu</td>
<td></td>
</tr>
<tr>
<td>Learning with Conversational AI and Personas: A Systematic Literature Review</td>
<td>1003</td>
</tr>
<tr>
<td>Antun Drobnjak, Ivica Boticki, Peter Seow and Ken Kahn</td>
<td></td>
</tr>
<tr>
<td>Supporting Content Creators in Creating Accessible Digital Materials in Higher Education</td>
<td>1009</td>
</tr>
<tr>
<td>Eirik Hansen and Weiqin Chen</td>
<td></td>
</tr>
</tbody>
</table>
AIED/ITS
An Efficient and Generic Method for Interpreting Deep Learning based Knowledge Tracing Models

Deliang WANG\textsuperscript{a,c}, Yu Lu\textsuperscript{a,b}*, Zhi ZHANG\textsuperscript{a} & Penghe CHEN\textsuperscript{a,b}

\textsuperscript{a}School of Educational Technology, Faculty of Education, Beijing Normal University, China
\textsuperscript{b}Advanced Innovation Center for Future Education, Beijing Normal University, China
\textsuperscript{c}Faculty of Education, The University of Hong Kong, Hong Kong, China

\textsuperscript{*}luyu@bnu.edu.cn

Abstract: Deep learning-based knowledge tracing (DLKT) models have been regarded as the promising solution to estimate learners' knowledge states and predict their future performance based on historical exercise records. However, the increasing complexity and diversity make DLKT models still difficult for users, typically including both learners and teachers, to understand models' estimation results, directly hindering the model's deployment and application. Previous studies have explored using methods from explainable artificial intelligence (xAI) to interpret DLKT models, but the methods have been limited in their generalizing capability and inefficient interpreting procedures. To address these limitations, we proposed a simple but efficient model-agnostic interpreting method, called Gradient\textsuperscript{*}Input, to explain the predictions made by these models in two datasets. Comprehensive experiments have been conducted on the existing five DLKT models with representative neural network architectures. The experiment results showed that the method was effective in explaining the predictions of DLKT models. Further analysis of the interpreting results revealed that all five DLKT models share a similar rule in predicting learners' item responses, and the role of skill and temporal information was found and discussed. We also suggested potential avenues for investigating the interpretability of DLKT models.

Keywords: Knowledge tracing models, deep learning, explainable artificial intelligence

1. Introduction

The ability to automatically identify learners' knowledge states is crucial for personalized learning, and it plays a fundamental role in sustaining learning motivation (Pelánek et al., 2017) and improving academic performance (Koedinger & Aleven, 2016). To achieve this, researchers have developed a range of knowledge tracing (KT) models by leveraging on learners' historical exercise records to predict their future performance. With the advancement of artificial intelligence (AI), KT models that employ deep learning techniques are considered effective due to their strong capability to capture inherent information. However, the complex structures and large number of variables in the deep learning-based knowledge tracing (DLKT) models make them difficult for users (e.g., teachers, students, and education researchers) to understand the models' decisions (Tsai & Gasevic, 2017), which may reduce users' trust and accordingly hinder the DLKT models' deployment and application, as indicated in the case of automated recommendation systems (Dietvorst et al., 2015). Additionally, blindly trusting the incorrect decisions would cause a wrong diagnosis of knowledge status and accordingly reduce learning efficiency.

To address the interpretability issue in deep learning knowledge tracing (DLKT) models, researchers have started exploring solutions. One approach is the incorporation component of the model remains a black box. Another approach is using the explainable artificial intelligence (xAI) techniques to interpret DLKT models, such as the model-specific layer-wise relevance propagation (LRP) method (Lu et al., 2020; Lu et al., 2022), but it is
only applicable to the specific DLKT models and hard to be generalized to other DLKT models. The DeepSHAP method has been also proposed as a more generic method for interpreting DLKT models (Wang et al., 2022), but it heavily relies on reference samples, which are normally scarce in the real cases. In addition, it is still in lack of systematic study on interpreting the existing representative DLKT models, typically including the recurrent neural networks (RNNs), memory-augmented neural networks (MANNs), attention, graph neural networks (GNNs), and convolutional neural networks (CNNs).

We thus propose a cost-effective and generic method to interpret the diverse DLKT models, where the model-agnostic interpreting method is designed to explain the predictions made by the five representative DLKT models. The comprehensive experiments have validated the effectiveness of the proposed method. Further investigations show that despite their different configurations, the five DLKT models follow a similar rule for estimating learners’ knowledge states and making predictions.

2. Related Work

2.1 Deep Learning based Knowledge Tracing (DLKT) Models

DKT (Piech et al., 2015) was the pioneer DLKT model, and it utilized RNNs to achieve superior performance compared to traditional Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994). Researchers subsequently improved upon the DKT model by incorporating additional input features such as question difficulty (Sonkar et al., 2020) and extending the model’s structure such as adding another layer of RNN (Su et al., 2018). Other types of deep neural networks have also been adopted for use in DLKT models. For example, because DKT summarizes the states of all concepts in a single vector, memory-augmented neural networks (MANNs) were used to build the DKVMN (Zhang et al., 2017), SKVMN (Abdelrahman & Wang, 2019), and EKT (Liu et al., 2019) models, which store the status of each concept in a separate space. Due to the limited performance of RNN-based and MANN-based KT models on concepts with little data, the attention mechanism has been employed to extract the similarity between different concepts and questions, leading to the development of the SAKT (Pandey & Karypis, 2019) and AKT (Ghosh et al., 2020) models. To capture the interconnected relationship between concepts and questions, graph neural networks (GNNs) have been used to build relationship graphs, resulting in the creation of the GKT (Nakagawa et al., 2019) and GIKT (Y. Yang et al., 2020) models. Additionally, convolutional neural networks (CNNs) have been explored for use in DLKT models, resulting in the development of the CKT (Shen et al., 2020) and CAKT (S. Yang et al., 2020).

2.2 Interpreting Methods for Deep Learning Models

Researchers have developed multiple interpreting methods to understand the internal workings and individual decisions of non-transparent deep learning models. These interpreting methods can be classified into model-specific and model-agnostic approaches (Adadi & Berrada, 2018). The former ones are tailored to the models with specific structures, while the latter ones can be applied to a wider range of models. Model-agnostic interpreting techniques include visualization (Goldstein et al., 2015), knowledge extraction (Hinton et al., 2015), influence methods (Cortez & Embrechts, 2011), and example-based explanations (Wachter et al., 2017). Influence methods are commonly used to explain models by estimating the importance or contribution of a feature to the model’s prediction. Popular methods in this category include perturbation-based and backpropagation-based approaches, such as Gradient (Simonyan et al., 2013) and DeepSHAP (Lundberg & Lee, 2017).

Although there are a large number of interpreting methods in the field of explainable AI (xAI), currently only two methods (i.e., LRP and DeepSHAP) have been utilized to interpret DLKT models (Lu et al., 2020; Wang et al., 2022). These methods are limited in their generalizing capability or their high computational complexity, leading to the need for simpler yet effective methods to provide explanations of the diverse and complex DLKT models.
3. Interpreting Method

3.1 Method

We propose to utilize an existing xAI method, called Gradient*Input (Arras et al., 2019), for interpreting DLKT models. Compared to other methods previously investigated for DLKT models (i.e., LRP and DeepSHAP), Gradient*Input has two advantages. First, it can be applied to a wide range of models regardless of their internal structure. Second, it has relatively low computational complexity, requiring only one forward pass and one backward pass to obtain an explanation, and does not rely on auxiliary information, such as the reference samples used in DeepSHAP. And this method has been demonstrated to be effective in various tasks (Arras et al., 2019; Shrikumar et al., 2017).

The Gradient*Input method explains the model prediction by decomposing it into the contributions of the input features. Formally, given an input \( x \) with \( K \) features, the prediction of a deep learning model on class \( c \) is a highly non-linear function \( f_c(x) \). Based on Taylor expansion (Li et al., 2016; Montavon et al., 2017), the non-linear prediction \( f_c(x) \) is approximated by a linear function, as shown in Equation 1:

\[
    f_c(x) \approx \sum_{i=1}^{K} \frac{\partial f_c(x)}{\partial x_i} x_i + b
\]

where \( i \) represents the \( i \)-th feature in the input, \( \sum_{i=1}^{K} \frac{\partial f_c(x)}{\partial x_i} \) represents the partial derivative of \( f_c(x) \) with respect to \( x_i \), and \( x_i \) reflects the contribution of feature \( x_i \) to the prediction.

![Figure 1. The interpreting procedure of Gradient*Input for DLKT models.](image)

3.2 Explaining DLKT Models

Figure 1 illustrates the procedure of using Gradient*Input to interpret the predictions of DLKT models. Specifically, given a set of question-answer records from learners, denoted as \( \{(q_1, a_1), (q_2, a_2), ..., (q_n, a_n)\} \), DLKT models can make predictions about their future performance, e.g., the probability \( f(n+1) \) of correctly answering a future question \( q_{n+1} \), as Figure 1(a) shows. To explain the prediction, Gradient*Input uses backpropagation to calculate the partial derivative of the prediction with respect to the input features, as depicted in Figure 1(b). Then, the element-wise product between the feature and its derivative is performed, which allows it to obtain the feature relevance (i.e., contributions) for each question-answer record, as Figure 1(c) indicates. Because each record is represented as a vector (e.g., an embedding or a one-hot vector with \( m \) dimensions), the sum of the feature relevance in a question-answer record can be used to determine the overall relevance of the record to the prediction (i.e., QA relevance), as shown in Figure 1(d).
4. Evaluation

4.1 Construction of DLKT Models

We select five typical and representative existing DLKT models for the experiment, including (1) DKT (Piech et al., 2015), the first DLKT model that adopts RNN as its inner structure; (2) DKVMN (Zhang et al., 2017), a DLKT model that adopts MANN to store concept status; (3) AKT (Ghosh et al., 2020), a DLKT model that employs attention to extract similarity between concepts and questions; (4) GKT (Nakagawa et al., 2019), A DLKT model that uses GNN to build skill relationship graphs; (5) CKT (Shen et al., 2020), A DLKT model that utilizes CNN to model individualization.

4.1.1 Dataset

We adopt two commonly used KT datasets, ASSISTment2009 and ASSISTment2015 (Feng et al., 2009). Specifically, the datasets were preprocessed to eliminate repetitive data and data without skill or question tags, and the length of the learner answer sequence was set to range from 10 to 200. After preprocessing, ASSISTment2009 contained 325,637 records on 110 skills and ASSISTment2015 contained 682,223 records on 100 skills. 80% of the data was randomly chosen for training, while the remaining 20% was used for testing.

4.1.2 Model Training

For all five models, we uniformly set the optimizer, dropout rate, mini-batch size, initial learning rate, and iteration epoch to Adam, 0.5, 64, 0.005, and 100, respectively. For the DKT model, the hidden dimension was set to 64. For the DKVMN model, the state dimension was set to 64 and the memory size was set to 110 for the ASSISTment2009 dataset and 100 for the ASSISTment2015 dataset. For the AKT model, the hidden dimension was set to 256 and the number of heads was set to 8. For the GKT model, the hidden dimension was set to 64 and the number of heads was set to 4. For the CKT model, the hidden dimension was set to 64. The performance (i.e., AUC and accuracy) of these five DLKT models in ASSISTment2009 and ASSISTment2015 can be seen in Table 1. Note that given the focus of this work is not on the model performance, we do not optimize the accuracy for each model.

### Table 1. The performance of five DLKT models in ASSISTment2009 and ASSISTment2015.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>ASSISTment2009</th>
<th>ASSISTment2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-based DKT</td>
<td>AUC</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>MANN-based DKVMN</td>
<td>AUC</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>Attention-based AKT</td>
<td>AUC</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>GNN-based GKT</td>
<td>AUC</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>CNN-based CKT</td>
<td>AUC</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>0.72</td>
<td>0.74</td>
</tr>
</tbody>
</table>

4.2 Method Validation

We first validate the capability of the proposed method on interpreting the decisions of DLKT models. We split the test data for both ASSISTment2009 and ASSISTment2015 into sequences of 15 question-answer records, resulting in 48,670 sequences for ASSISTment2009 and 97,637 sequences for ASSISTment2015. In each sequence, the first
14 records were used as input to predict the correctness of the last record, allowing us to identify correctly-predicted sequences. The number of positive and negative predictions made by all five DLKT models in each dataset is shown in Table 2. By applying the proposed method, we could calculate the relevance of each question-answer record to the prediction.

Table 2. The number of positive and negative predictions in correctly-predicted sequences.

<table>
<thead>
<tr>
<th>DLKT Models</th>
<th>ASSISTment2009</th>
<th>ASSISTment2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>RNN-based DKT</td>
<td>26,258</td>
<td>8,777</td>
</tr>
<tr>
<td>MANN-based DKVMN</td>
<td>27,477</td>
<td>7,660</td>
</tr>
<tr>
<td>Attention-based AKT</td>
<td>27,271</td>
<td>7,628</td>
</tr>
<tr>
<td>GNN-based GKT</td>
<td>27,005</td>
<td>8,066</td>
</tr>
<tr>
<td>CNN-based CKT</td>
<td>26,629</td>
<td>8,679</td>
</tr>
</tbody>
</table>

Figure 2. Question-answer record deletion results for DLKT models in ASSISTment2009.

Figure 3. Question-answer record deletion results for DLKT models in ASSISTment2015.

To validate the effectiveness of the relevance of each question-answer record, we further conducted the experiment by removing question-answer records with high relevance from each sequence and observing the resulting change in accuracy. Specifically, we deleted question-answer records in the descending order of relevance for positive predictions (i.e., predictions of a correct response in the last question), and in the ascending order of relevance for negative predictions (i.e., predictions of a wrong response in the last question). By observing the change in accuracy, we could evaluate whether question-answer
records with high relevance were more important for model predictions. We also performed random removal of records for comparison.

Figure 2 and 3 show the results of deleting question-answer records for all five DLKT models in the ASSISTment2009 and ASSISTment2015 datasets, where DLKT-GI represents the proposed method and DLKT-R means random deletion. As can be seen, compared to random deletion, the proposed method leads to a significant drop in accuracy. For example, in the case of the DKT model, removing 5 question-answer records based on relevance causes the prediction accuracy to drop from 100% to around 57%, while random deletion only slightly reduces the accuracy to around 95%. Similarly, for other DLKT models, deletions based on relevance cause the accuracy to drop to a range between 65% and 71%, while random deletions only reduce the accuracy to around 95%. Overall, the results suggest that the proposed method is effective in explaining the decisions of all five DLKT models.

4.3 Interpreting Model Rules

With the validated QA relevance, we further interpret how the DLKT models make predictions. We mainly analyze how the model input question-answer records influence the model output, i.e., the predictions on learners’ future performance. In particular, we consider the effects of both the skill information (i.e., the specific skill being tested) and temporal information (i.e., when the exercise was completed) on the model’s decision.

We selected correctly-predicted sequences and set the skill and position of the last record in each sequence as the target skill and prediction position. Then each question-answer record in the sequence was tagged as either Same skill or Different skill based on whether its skill matched the target skill, and Recent (i.e., the first half QA records) and Distant (i.e., the second half QA records) based on its distance from the prediction position. This resulted in four groups (i.e., 2 × 2): Distant & Same skill, Distant & Different skill, Recent & Same skill, and Recent & Different skill. For each group, the absolute relevance of each question-answer record was summed to compute the mean. We computed and compared the mean of four groups among all correctly-predicted sequences across the five DLKT models in the ASSISTment2009 and ASSISTment2015 datasets.

![Figure 4](image1)

Figure 4. Skill-distance relevance comparison for RNN-based DKT.
Figure 5. Skill-distance relevance comparison for MANN-based DKVMN.

![Graph](image1)

(a) ASSISTment2009  
(b) ASSISTment2015

Figure 6. Skill-distance relevance comparison for Attention-based AKT.

![Graph](image2)

(a) ASSISTment2009  
(b) ASSISTment2015

Figure 7. Skill-distance relevance comparison for GNN-based GKT.

![Graph](image3)

(a) ASSISTment2009  
(b) ASSISTment2015

Figure 8. Skill-distance relevance comparison for CNN-based CKT.

![Graph](image4)

(a) ASSISTment2009  
(b) ASSISTment2015

The results, shown in Figure 4, 5, 6, 7, and 8, indicate that for all five DLKT models, the recent-same-skill records (i.e., records that are close to the prediction position and on the target skill) have the highest relevance to the prediction, while those distant-different-skill records (i.e., records that are far from the prediction position and on non-target skills) have the lowest relevance. For example, in the case of AKT, the relevance of recent-same-skill records is approximately 0.036 in both the ASSISTment2009 and ASSISTment2015 datasets, while the relevance of distant-different-skill records is only about 0.005. The other four DLKT models show similar results for these two groups. Furthermore, it is difficult to differentiate between the relevance of recent-different-skill records or distant-same-skill records, as these groups show similar relevance in some cases (e.g., DKT and CKT) and differing relevance in others (e.g., DKVMN in ASSISTment2009 and ASSISTment2015).

We also find that the same skill records are more sensitive to distance (i.e., time) compared to different skill records. Specifically, for all five DLKT models in both ASSISTment2009 and ASSISTment2015, the average relevance difference between recent-same-skill records and distant-same-skill records is much larger than that between recent-
different skill records and distant-different-skill records. In contrast, distance seems to have a small effect on the relevance of different skill records for model prediction. For instance, the relevance of recent-same-skill records for the DKT model in ASSISTment2015 is about 0.055, 0.045 higher than distant-same-skill records, while the difference between recent-different-skill records and distant-different-skill records is close to 0.002. Other DLKT models show similar results.

Based on the findings discussed above, we conducted an additional experiment to further investigate the influence of skill and distance information on the decisions of DLKT models. Specifically, for each input sequence of length 15, we excluded the records that were distant from the prediction position and on non-target skills (i.e., distant-different-skill records) and thus only kept the records that were close to the prediction position or on the target skill (i.e., recent-same-skill records, recent-different-skill records, and distant-same-skill records). The leftover records made up 73% of the raw records in the ASSISTment2009 dataset and 63% in the ASSISTment2015 dataset. We then compared the model prediction performance with cases using all the raw records.

Figure 9 illustrates the experiment results: we see that for all five DLKT models, despite containing significantly fewer data points, the models with the leftover records have similar prediction accuracy to the models with raw full sequences. For example, DKVMN achieves 0.734 in accuracy with 1,464,555 records in ASSISTment2015. After excluding about 37% of data, the accuracy of DKVMN remains at 0.734 with 932,799 records. The experiment results partially validate the findings and the rules that how the DLKT models utilize the skill and distance information to make the decision.

![Figure 9](image)

**Figure 9.** The influence of skill and distance on the decisions of DLKT models.

5. Conclusion

In this work, we propose a simple and efficient xAI method to address the interpretability issue of increasingly complex and diverse DLKT models. The method only requires a lower complexity and can be used to explain a wide range of DLKT models. The experiment results on five DLKT models in two datasets validate the effectiveness of the proposed method. Further analysis of the explanations reveals that all five DLKT models use a similar rule when making the decision: question-answer records that are close to the predicted question and on the same skill as the predicted question are found to be the most relevant indicators, while records that are distant and on different skills are the least important. We also have found that records of the same skill are more sensitive to changes in distance compared to records of different skills. Additionally, it is also observed that using fewer but relevant question-answer records to make predictions can achieve similar accuracy as using full sequences, which supports the findings about the decision rules of DLKT models.

This work has significant impacts on practice. First, the rules obtained can be integrated into intelligent tutoring systems that utilize DLKT models. Automatically identifying learners' knowledge states and providing explanations can potentially increase their trust in the system and help them adjust their learning behavior when receiving incorrect diagnoses of their knowledge states. Second, the findings about the decisions of DLKT models...
contribute to making these models more transparent and provide valuable insights for researchers to design more interpretable KT models. It can be a promising direction for future research to evaluate the impact of these explanations on education and consider the effect of skill and temporal information when designing DLKT models.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (no. 62077006, 62177009) and in part by the Fundamental Research Funds for the Central Universities.

References


Abstract: Many studies applying neural networks to the field of education have focused on student performance prediction and explainability of their decisions. While those studies introduced neural networks into educational settings, such networks cannot directly support student learnings in place of teachers. Therefore, we present a method that uses a general Transformer encoder to recommend appropriate learning actions for improving student performance. By considering the attention weight of a low-performing student to be close to that of a high-performing student, our method recommends the learning materials and actions for learning the materials. To evaluate the effectiveness of our method, we trained a deep neural network (DNN) on a private dataset of student operations (e.g., NEXT, PREV, OPEN) on digital learning materials obtained from a Japanese university. The number of operations divided by each learning material and by type of operation are input to the DNN, and the DNN outputs the student’s grade on 5-point scale. We applied our method with this trained DNN to samples that successfully predicted grades, and the number of operations increased on the basis of the recommended learning materials and actions. By re-inputting modified sample into the DNN, we then observe how the student performance changes. The results of this simple experiment indicate that more students improved their performance with both the material-based and operation-based recommendations than with random recommendations. The percentage of students whose grades improved tended to be larger for those with low grades. Specifically, the improvement ratio for students with the two lowest grades was over 90% by operation-based recommendation. This is consistent with our intuition that low-performing students are more likely to improve.

Keywords: Student Performance Prediction, Explainable AI, Transformer

1. Introduction

Deep neural networks (DNNs) have been actively studied in various fields such as image processing and natural language processing. They are expected to be used for education in the current remote educational environment triggered by the COVID-19 pandemic (Adedoyin & Soykan, 2020). In an educational environment where there is physical distance between
teachers and students, it is difficult for teachers to provide individual feedback, so support from DNNs is necessary.

There have been many studies introducing DNNs into educational environments (Piech et al., 2015; Imran et al., 2019; Abdelrahman & Wang, 2019; Xing, & Du, 2019). These studies showed that we can use DNNs to predict a student's performance and possibility of early school dropout. However, it is necessary to incorporate explainability and accountability in educational DNNs design (Webb et al., 2021). There many studies explained the basis for DNN predictions (Baranyi et al., 2020; Mu et al., 2020; Vultureanu-Albiş & Bădică, 2021; Hasib et al., 2022; Swamy et al., 2022). However, these studies did not focus on DNNs replacing teachers, thus could not directly reduce the burden on teachers.

We propose a method with which artificial intelligence (AI) instead of teachers recommend appropriate learning actions (e.g., selection of learning materials, operations on learning materials) for improving student performance (see Section 2). We developed this method to misidentify the output of AI predicting student grades as a good grade by minimally perturbing the input of a student with a low grade. The pipeline for integrating our method into an educational environment is shown in Figure 1. We conducted an experiment to evaluate our method, and the results are presented in Section 3. Here, the predicted grades are denoted by S (excellent) , A (good), B (satisfactory), C (pass), and F (fail).

![Figure 1. Pipeline for Deploying DNN into an Educational Environment.](image)

2. Proposed Method

Given a DNN and dataset including student learning actions on digital-learning materials, we developed a method for recommending learning actions to improve grades in an input-data dependent manner. We first introduce the criterion to identify learning actions.

2.1 Identification of Learning Actions that Significantly Affect Predictions

When an input vector \( x \) is perturbed \( \delta \), the network loss \( \mathcal{L} \) changes as:

\[
\Delta \mathcal{L}(x) = |\mathcal{L}(x) - \mathcal{L}(\delta \odot x)|, \tag{1}
\]

where \( \odot \) denotes the Hadamard product. Next, to compute the impact of each input, Equation (1) is transformed with first order Taylor expansion by focusing on a single element \( x_i \in x \):

\[
\Delta \mathcal{L}(x_i) = |\mathcal{L}(x_i) - \mathcal{L}(\delta x_i)|
= \left| \mathcal{L}(x_i) - \mathcal{L}(x_i) - \frac{\partial \mathcal{L}}{\partial x_i} (\delta x_i - x_i) - \mathcal{O}(\|x_i\|^2) \right|
\approx \left| \frac{\partial \mathcal{L}}{\partial x_i} (\delta x_i - x_i) \right|, \tag{2}
\]

where \( \mathcal{O} \) denotes terms of higher order than the 2nd degree. We evaluate the input data on material or operation in accordance with Equation (2).
Materials: If the number of learning materials is $M$ and the type of operations (e.g., NEXT, PREV, OPEN) is $N$, the input data are denoted as $X \in \mathbb{R}^{M \times N}$. When $X$ is input, the score of material $m$ is expressed by using Equation (2) as the following equation:

$$S_m(X) = \sum_{n=1}^{N} \left| \frac{\partial L}{\partial x_{m,n}} (\delta x_{m,n} - x_{m,n}) \right|,$$

where $x_{m,n}$ represents the number of operations per type of operation in $m$. We select the material with the highest score and recommend it to students.

Operations: The score for each operation is expressed as:

$$S_n(X) = \sum_{m=1}^{M} \left| \frac{\partial L}{\partial x_{m,n}} (\delta x_{m,n} - x_{m,n}) \right|.$$

We identify the type of operation that is effective for learning by using Equation (4). As with the material criterion, we recommend to students the operation with the highest score.

2.2 Optimization for Transformer

The theory introduced in Section 2.1 is applied independently of DNN structure. However, calculation of automatic differentiation using the loss between the correct label and output incurs large computational cost. Therefore, we introduce an additional lightweight criterion optimized for a Transformer (Vaswani et al., 2017).

A Transformer introduces an attention mechanism to determine the dependencies of various ranges (e.g., shorter, or longer range) within a sequence. In this attention mechanism, attention weight $A$ in the first layer is derived using input $X \in \mathbb{R}^{M \times M}$ as:

$$A = \text{Softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) \in \mathbb{R}^{M \times M},$$

s.t. $Q = XW_q, K = XW_k$,

where $d_k$ is a scaling coefficient, and $W \in \mathbb{R}^{d \times M}$ denotes learnable parameters. By considering the interrelationships between input sequences using Equation (5), we can determine which inputs are contributing to the prediction.

Given an ideal attention weight $A'$ of a high-performing student, we can reduce the computational cost from Equation (3) and (4). The method is simply to replace the loss in Equations (3) and (4) by the squared loss between $A'$ and $A$:

$$\mathcal{L} = \sum_{j=1}^{M} \sum_{i=1}^{M} (a'_{i,j} - a_{i,j})^2,$$

s.t. $a'_{i,j} \in A'$, $a_{i,j} \in A$.

This optimization method makes it possible to recommend learning actions only through processing by the first layer of Transformer encoders.

3. Experiment

We evaluated the recommendation of learning actions on a private dataset obtained in a real educational environment. To verify that the ideal attention weight is obtained from the training data, we first investigate the attention weight per grade. We then investigated the validity of recommending learning actions.

3.1 Dataset

We use log data of real student learning actions and the grades collected at a Japanese university. These log data were compiled for 21 types of operations (e.g., NEXT, PREV, OPEN) executed on 12 types of materials in a single course. The number of training data.
was 114, and the number of test data was 51, because we used one year of such data for evaluation collected for a period of three years 2020, 2021 and 2022.

Figure 2 illustrates how the data were preprocessed. We treated the number of operations for each type of operation as a histogram for each learning material. Thus, the input for a student was $X \in \mathbb{R}^{12 \times 21}$. The values of these histograms were normalized by the maximum number of operations calculated using the histograms of all students.

Figure 2. Preprocessing of The Data of One Student.

### 3.2 Experimental Setup

We construct a success-prediction model that can accurately classify the corrected label (final grade) $y_s$ for student $s$ from learning actions, as shown in Figure 2. We rely on a DNN architecture that is based on Transformer encoders, which can combine high with reduced computational cost due to Equation (6). An overview of the model structure is shown in Figure 3. The features input to the model are linear projections from student data such in Figure 2. The number of Transformer encoders $L$ was 4. The model was trained using Adam (Kingma & Ba, 2015) for 200 epochs with an initial learning rate of 0.0001 and batch size of 32. The loss function uses cross-entropy loss.

Figure 3. Overview of Model Structure for Grade Prediction.

### 3.3 Prediction of Student Grades

To coordinate the number of embedding dimensions $d$ and heads $h$, we compare
their prediction accuracy with Transformer models of various scales. Table 1 shows the top-1 accuracy of grades on the test dataset. The number of embedding dimensions was investigated in the range from 64 to 256, and the highest test accuracy was achieved at 128. The highest accuracy was then obtained with one head in all comparisons. Using the results of this experiment as a guide, we recommend learning actions using a model with 128 embedding dimensions and one head.

3.4 Attention Weight Differences Between Grades

<table>
<thead>
<tr>
<th>$d$</th>
<th>$h$</th>
<th>Top-1 Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>1</td>
<td>39.21</td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>39.21</td>
</tr>
<tr>
<td>128</td>
<td>1</td>
<td>52.94</td>
</tr>
<tr>
<td>128</td>
<td>2</td>
<td>47.06</td>
</tr>
<tr>
<td>256</td>
<td>1</td>
<td>45.10</td>
</tr>
<tr>
<td>256</td>
<td>2</td>
<td>45.10</td>
</tr>
<tr>
<td>256</td>
<td>4</td>
<td>27.12</td>
</tr>
</tbody>
</table>

Attention is conducted multiple times in parallel, which results in higher performance than just using a single head. However, our results were the exact opposite (see Section 3.3). This means that mixing features between materials does not require such complex patterns. We hypothesize that if the mixing between materials is represented by a simple pattern, there is an ideal attention weight, as described in Section 2.1, common to high-performing students. To confirm this hypothesis, we investigated the attention weight by grade.

Figure 4 shows the attention weights of the first layer, calculated using the training dataset. In Figure 4, high values of attention weight are shown in light colors and low values in dark colors. These attention weights are averaged per grade. Since the number of teaching materials was 12, the size of attention weight was 12x12. We observed that the attention weights of the high-performing students, such as those with grades S and A, had a biased attention to fewer materials. The attention of the lowest-performing students are plotted across many learning materials. Therefore, we use the attention weight of students with grade S as an ideal attention weight to reduce the computational cost of recommending learning actions.

![Figure 4. Average of Attention Weight per Grade Obtained from First Layer of Transformer Encoder.](image)
3.5 Recommending Learning Actions

Our method recommends appropriate materials and operations to a student using
the trained DNN.

**Material Recommendation:** Our method recommends the materials that students
should be learning to improve their grades. To investigate such improvement, we
increased the number of operations included in recommended materials by a factor
of 1.5. We then re-input the sample with an increased number of operations into the
DNN to check for changes in performance. Figure 5 shows the ratio of performance
improvement for the sample that successfully predicted grade before modification
when our method was used to recommend materials to be learned. For comparison,
Figure 5 plots the improvement ratio with dashed lines when learning materials were
randomly recommended (the results plotted with solid lines are from our method).
The vertical axis is the ratio of predicted improved performance after re-input, and
the horizontal axis is the number of materials that were modified using our method or
randomly.

High rate of improvement was observed in the lowest-performing students
with grades C and F. Their performance tended to improve more in proportion to the
number of materials modified. These results are consistent with the intuition that
lower-performing students are more likely to improve with a small amount of
learning, and that students who learn more improve their performance. Then, the
improvement ratio with our method is higher than that with random, which confirms
the effectiveness of our method.

![Figure 5. Ratio of Performance Improvement for Each Material Modified.](image)

**Operation Recommendation:** We then identified the types of operations that
should be used to improve student performance. By modifying the number of
operations by a factor of 1.5 for our method, or for a randomly recommended
operation, we can compare the changes in grade predictions. Figure 6 shows the
ratio of performance improvement for the sample that successfully predicted grade
before modification when our method was used to recommend operations to be
learned. The vertical axis is the percentage of predicted grades that improved after
re-input, and the horizontal axis is the number of operations that were modified using our method or randomly.

![Figure 6](image)

Figure 6. Ratio of Performance Improvement for Each Operation Modified.

The operations recommended with our method were more accurate than those randomly recommended. From these results, the performance of students with low grades improved significantly with our method.

4. Discussion

Our results showed that the suggested learning actions improved student performance. However, our method is highly dependent on the quality and quantity of the learning data and may not work well when the number of students is small or when there are not enough learning logs (our method may be very accurate in the presence of a large number of learning logs). Figure 7 shows the top-5 most frequently recommended operations and their average scores. Many basic operations (e.g., NEXT, PREV, OPEN) are recommended, and operations such as “ADD MEMO” are not included. While the basic operations are performed many times and are understandably effective for estimating student learning density, most realistic teachers would recommend leaving notes and bookmarking important passages. Training on a larger dataset is necessary to achieve the same performance as a teacher using a DNN. Therefore, a larger dataset would be needed to construct a DNN comparable with a teacher. Since suggesting learning actions using too small a dataset may invade the privacy of certain highest-performing students, one should be careful when deploying a pipeline such as that shown in Figure 1.
5. Conclusion

We showed that considering the impact of inputs on grade prediction can recommend the learning actions needed to improve grades. Learning actions are recommended on a per-material or per-operation basis, allowing the DNN to provide detailed feedback to the students. In particular, the operation recommendations confirmed a higher improved ratio of grades than the learning material recommendations. Recommending learning actions to students is a necessary component for neural networks to directly assist teachers and encourage the introduction of AI into the educational environment. In future work, we plan to use data augmentations suitable for event logs to recommend learning actions that are more in line with our intuition.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR22D1, Japan.

References


An agent-based modeling and simulation tool as a learning aid for diffusion of innovations

Joseph Benjamin ILAGAN\textsuperscript{a}, Jose Ramon ILAGAN\textsuperscript{a} \& Maria Mercedes RODRIGO\textsuperscript{a}
\textsuperscript{a}Ateneo de Manila University, Philippines
\textsuperscript{a}jlilagan@ateneo.edu

Abstract: The diffusion of innovations describes how new technologies spread through a population. Agent-based modeling and simulation (ABMS) cover interactions among autonomous agents and the analysis of emergent outcomes from the behaviors, reactions, and interactions of these agents. Existing studies using ABMS aim to illustrate dynamic agent behavior and interaction with other agents and the environment in the context of diffusion of innovations. However, they do not extend the use of the simulators to teaching and learning. This study describes an ABMS built using NetLogo that allows students to explore the impact of various agent characteristics, behaviors, and interactions on adopting new innovations. Students can manipulate certain parameters involving the characteristics of these agents, such as their level of innovation propensity, social influence, and connectivity, to see how these factors influence the adoption of the innovation. The outcome of each run is logged, analyzed, and presented to the students as meaningful feedback and suggestions for supplementary learning from an LMS in preparation for succeeding simulation iterations. The simulator preserves agent autonomy and adaptability while allowing the students to play with model parameters. As the simulator transitions to more empirical data for rules governing the behaviors of agents, future versions of the simulation may incorporate additional user interface and AI-based simulator elements.

Keywords: agent-based modeling, agent-based model simulation, computer simulation, entrepreneurship education, diffusion of innovations

1. Introduction

The diffusion of innovations describes how new ideas and technologies spread through a population (Rogers, 1983). In the context of a complex social system, the diffusion of innovations theory reflects adoption timing and decisions made by consumer groups classified based on the following adoption timing classification theme: (1) innovators, (2) early adopters, (3) early majority, (4) later majority, and (5) laggards (Rogers, 1983). Diffusion of innovations is relevant in entrepreneurship education (EE). However, traditional teaching approaches to entrepreneurship have been based on business management education, which is inflexible and ineffective (Fayolle & Gailly, 2008). Computer simulation (Roberts, 2011) helps develop, teach, and test entrepreneurship theory and helps entrepreneurs understand, predict, and explain the potential consequences of their decisions. Agent-based modeling and simulation (ABMS), a form of computer simulation, covers interactions among independent agents (Macal & North, 2008) and the analysis of emergent outcomes from the behaviors and interactions of these agents (Garcia, 2005). This paper is part of a larger set of studies using agent-based modeling and simulation for EE (Ilagan, 2022). The focus is on diffusion of innovations as a simulation scenario to understand the mechanisms that drive the adoption and spread of new ideas and technologies. The study also addresses the potential for integrating artificial intelligence at the agent level. How should the simulator allow students to play around with different characteristics of agents, impacting the adoption of innovations in the simulation while preserving the agents' autonomy to allow for emergent
outcomes? RQ1) How can autonomy and adaptability of agents be implemented while giving students the freedom to experiment with various model parameters? RQ2) What software architecture is needed for the ABM simulation to have ample validity for learning the concepts of diffusion of innovations? RQ3) In what ways can the ABM simulation support Artificial Intelligence at the agent level?

2. Related work

With the diffusion of innovations as the simulation scenario, students may experiment with various technology strategies: from pursuing technology leadership as an innovator to behaving as an early adopter or a fast follower. They can experience the respective advantages and drawbacks of these strategies (Stummer & Kiesling, 2021). Specific skills imparted by the simulation environment related to diffusion of innovations, as inspired by Stummer & Kiesling (2021), are market entry, global product launch strategies, pricing strategies, product lifecycle and portfolio management, and technology strategy. NetLogo (Tisue & Wilensky, 2004) is an open-source programming language and modeling environment specifically designed for agent-based modeling. While several studies have already featured diffusion of innovations implemented in ABMS, none have been explicitly tied to teaching or learning entrepreneurship. ABMS is argued to be either a subfield (Distributed AI) (O’Sullivan & Haklay, 2000) or an alternative of AI that takes advantage of the resulting emergence of interactions and the stochastic representation of intelligence (Luger, 2005). Machine learning (ML) and AI techniques may also enhance the behavior and capabilities of agents, as proposed by Rand (2006).

3. Approach

The parameters of the agent model representing different sensitivities or indexes (features, price, promotion, and social) influencing innovation adoption are based on the work of Schramm et al. (2010). The rules and formulas used in this study are simplified versions of those prescribed by Bass (1980, 2004). These include the consumer adoption threshold (CAT), a function of features, price, promotion, social influence, etc. The product price tends to drop over time. The expected S-curve related to adoption and the diminishing prices shall be reflected graphically and animated over time. The only form of relationship is when an agent is physically near another. Integration with a Learning Management System (LMS) can be done through file transfer of NetLogo logs or API callouts for content from LMS, where each simulation run is logged, analyzed, and presented to the students as meaningful feedback. Figure 1 shows the simulator’s state before and after the run. The layout of consumer agents resembles a bell curve. A chart on the right shows an S-curve for the adoption of the innovations, with all green indicating full saturation. Another chart shows revenue per period, indicating tapering off revenues as the innovation matures and eventually declines.

![Figure 1. Simulator state before and after the run.](image-url)
The Mobile switch allows agents to move around the screen to allow encounters with other agents and to influence their neighbors. Students can manipulate agents to observe how these factors influence the adoption of the innovation. Agents will behave autonomously based on the programmed rules, behaviors, and parameters set by the students.

4. Conclusion and Future Work

Future versions of this simulation will allow direct relationships and will not be constrained to proximity as the basis for being able to influence others. While this paper covers only rule-based methods for agent behavior, machine learning (ML) and AI techniques, in turn, will enhance the behavior and capabilities of agents (Rand, 2006) in future versions of this simulator. As the simulator transitions to more empirical data for rules governing the behaviors of agents, future versions of the simulation may incorporate additional user interface and AI-based simulator elements. In addition, rather than relying on formulas for consumer and agent behavior, external data fed into an ML model relied on by agents will allow more adaptive behavior in the simulator. Another shortcoming of this simulation is that the behavior of competitors is fixed and lacking. With the introduction of ML and AI techniques, adaptive competitor behavior will make the simulator more interesting to the technology entrepreneurship students as they will see consumer and producer (competitor) behavior. There also is an opportunity to take advantage of generative AI (GAI) and large language models (LLMs) to enable a conversational element to the educational interface.

References


Studying Memory Decay and Spacing within Knowledge Tracing

Cristina MAIER\textsuperscript{a}, Isha SLAVIN\textsuperscript{b}, Ryan S. BAKER\textsuperscript{c}, & Steve STALZER\textsuperscript{d}

\textsuperscript{a}McGraw-Hill Education, USA  
\textsuperscript{b}McGraw-Hill Education, USA  
\textsuperscript{c}University of Pennsylvania, USA  
\textsuperscript{d}McGraw-Hill Education, USA  
*cristina.maier@mheducation.com

Abstract: Knowledge Tracing estimates a student's knowledge on a set of skills and predicts whether the student will answer correctly if given a question linked to subsets of such skills. We conduct an in-depth analysis on the best ways to apply the cognitive science principles such as memory decay and spacing within knowledge tracing, proposing new algorithms, MemDec and MemDec Spacing, to do so. We explore different methods of modeling the rate and weight of decay, with and without spacing, and analyze their impact on predicting performance in real-world data. Variations of the model are compared between each other as well as against other existing algorithms.

Keywords: Knowledge tracing, memory decay, spacing effect, learning systems

1. Introduction

Knowledge tracing attempts to estimate student mastery on a set of skills from their performance. The outputs of knowledge tracing have several uses, including use in behavioral and self-regulated learning models, reports to instructors with actionable insights, and driving mastery learning within adaptive learning systems (Pelánek, 2017). Although the majority of the most recent work in knowledge tracing has investigated refinements to deep learning algorithms (Piech et al., 2015), there has also been interest in developing knowledge tracing algorithms that leverage findings from cognitive science (Choffin et al., 2019; Pavlik et al., 2021) and studying how broadly applicable they are (Schmucker & Mitchell, 2022). In this paper, we investigate the implementation of cognitive science principles within knowledge tracing in detail, specifically focusing on memory decay and the spacing effect when applied to variants of some components of Logistic Knowledge Tracing (LKT) (Pavlik et al., 2021) with an emphasis on the components from Performance Factor Analysis (PFA) (Pavlik et al., 2009) and Recent-Performance Factor Analysis (R-PFA) (Galyardt & Goldin, 2014). We chose to study memory decay and spacing within this framework because of PFA’s interpretability and its support of multi-skill items. To study these concepts, we introduce a new algorithm that incorporates memory decay into the knowledge estimation and investigate how the spacing effect and different ways of governing the increase in memory decay influence modeling student performance. We consider two methods for modeling memory decay: one that uses the practice order (as in Galyardt & Goldin, 2014), and a second that uses a time window approach. Incorporating these concepts allows for the differentiation between a student’s current knowledge versus previous comprehension that may have been lost over time.

In Section 2 we discuss the memory decay and spacing effect theories, and describe the well-known knowledge tracing algorithm, Performance Factor Analysis (PFA) (Pavlik et al., 2009). Section 3 contains Related Work investigating issues surrounding memory within knowledge tracing. In Section 4 we introduce a new algorithm, MemDec, and a variant, MemDec Spacing, that incorporates the spacing effect. Section 5 presents the real-world dataset used for the experiments. Analysis and experimental results are presented in Section 6. Finally, conclusions and final remarks are discussed in Section 7.
2. Knowledge Tracing and Memory

2.1 Memory Decay and Spacing Effect

Psychological effects influence student learning in classrooms and virtual learning environments. One such principle is the decay theory, a principle of forgetting which states that memory fades with the passage of time. Unless reinforced by repetition, the information we learn is forgotten over time. Without incorporating decay into knowledge estimation, it is difficult to account for situations where a student forgets. Memory decay has been widely studied and modeled by cognitive scientists and learning scientists (e.g. Mozer et al., 2009; Pavlik & Anderson, 2008). For example, in MCM, memory decay is incorporated through a power function in which each item-specific memory trace decays exponentially (Mozer et al., 2009). This model was then used to determine optimal practices to yield the highest retention of material in a classroom. Similarly, a study was conducted to measure and compare the relearning of forgotten material by three computational models, all of which incorporate a component of decay over time in its prediction equation (Walsh et al., 2018).

There is increasing evidence that knowledge retention can be enhanced through methods such as distributed learning, which sequences learning such that concepts are practiced in a distributed schedule over time rather than with massed schedules where practice attempts related to the same set of skills occur in quick succession. Research in this area has demonstrated that both time elapsed between practices of the same material and time elapsed between final study episode and an exam affect final-test retention (Wiseheart et al., 2006). Work has attempted to use models to determine optimal spacing, such as (Mozer et al., 2009; Pavlik & Anderson, 2008), each of which introduced a model that can predict the influence of specific study schedules on retention for specific items. Knowledge tracing algorithms incorporating spacing effects into learner models (e.g. Walsh et al., 2018; Pavlik et al., 2021) have been used for an increasing number of applications, including the adaptive sequencing of practice (Eglington & Pavlik, 2020). In total, there is considerable evidence that spacing is important for long-term retention of knowledge. Within this paper, we investigate different variations on how to represent the spacing effect in knowledge estimation, analyzing how these models perform for different representations of memory decay, through altering the rate and strength of decay over time.

2.2 Performance Factor Analysis

Performance Factor Analysis (PFA) is a knowledge tracing algorithm (Pavlik et al., 2009) that can be used with multi-skill items, in contrast to some of its extensions which only work for single-skill items (Galyardt & Goldin, 2014). PFA has been successful at predicting both knowledge within the learning system and latent knowledge carried outside the system (Scruggs et al., 2020). Our proposed algorithm and experiments use PFA as a baseline model.

The original PFA model predicts the performance of a student on a given item/problem, at a given time. It does this by using the student’s past number of successes, multiplied by a weight $\gamma$ fit to each of the item’s skills; a student’s past number of failures, multiplied by a weight $\rho$ fit to each of the item’s skills; a weight $\beta$ which represents the difficulty of an item. Depending on the variant of PFA, $\beta$ is either applied across all contexts, across all items linked to the current skill (the most common approach and what we will use here), across all items of the same “item-type”, or for individual items. These features are input into a logistic function to obtain a prediction, $p(m)$, which gives the probability of success for a given student on a given (future) item. The PFA formula is given below:

$$m(i; j \in KC; s; f) = \sum_{i \in KC} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

$$p(m) = \frac{1}{1 + e^{-m}}$$

where $i$ represents a student, $KC$ are the knowledge components (i.e. skills) linked to the item, $j$ represents a skill. Parameters $\beta_j$, $\gamma_j$ and $\rho_j$ are the learned parameters for skill $j$.

The PFA model in its original form does not incorporate the notion of memory decay. All previous practices are given the same weight, regardless of the time or order in which they
took place. As discussed above, some previous studies have incorporated the notion of memory decay into their knowledge tracing models. Doing so addresses the phenomenon that memory fades with the passage of time. Ignoring decay may be temporarily safe when practice sessions are massed (as in some intelligent tutoring systems) but will lead to less accurate inference when the student’s work on a skill is spread out over time.

3. Related Work

A variety of knowledge tracing frameworks and models have been proposed and studied. This includes methods based on Hidden Markov Models such as Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1995), which models a student’s latent knowledge of knowledge components as a set of binary variables which represent mastery or lack of mastery of each concept, and neural network models such as Deep Knowledge Tracing (Piech et al., 2015), which uses RNNs (recurrent neural networks) to learn concept patterns using long-short term memory (LSTM) without human annotations. There are also recent efforts to combine knowledge tracing and Item Response Theory (Oeda & Asai, 2016; Khajah et al., 2014b) with a decay effect, modeled as elapsed time and a forgetting parameter.

In the last few years, extensions to these categories of models have attempted to include item recency/decay. Researchers demonstrated that by adding a forgetting parameter to BKT, the algorithm is more sensitive to the effect of interspersed trials (Khajah et al., 2016). Another variant of BKT called Multistate BKT (Argawal et al., 2020) incrementally increases the weight of newer attempts. In Attentive Knowledge Tracing (Ghosh et al., 2020), researchers implemented a monotonic attention mechanism that uses an exponential decay curve to downweight past questions. DAS3H (Choffin et al., 2019) modeled both learning and forgetting curves, using factorization machines to handle multiple skills tagging. Learning Process-consistent Knowledge Tracing (LPKT) (Shen et al., 2021) models the student learning process as a set of tuples which includes the time series information of the assignments, thereby embedding both the answer time as well as the interval between activities.

Previous research has investigated variations of PFA that incorporate decay into their mastery predictions. One such model, PFA-Decay (Gong et al., 2011), took practice order into account using a decay factor $\delta$ ($0 < \delta \leq 1$) raised to a power representing the distance in practice number, and multiplied to the counts of successes and counts of failures. Another modification of PFA, Recent-PFA (R-PFA), incorporated memory decay into the model’s performance prediction in the form of a weighted proportion of success, with weight being dependent on the recency of the practice (Galyardt & Goldin, 2014). However, R-PFA does not take time specifically into account; it just considers practice order. R-PFA modifies the PFA formula by replacing the total number of failed practices, $f_{ij}$, with the total number of all practices thus far (essentially equating $f_{ij} + s_{ij}$), and replacing the total number of successful practices $s_{ij}$ with a component, $R_{ijt}$, that incorporates the notion of memory decay:

$$R_{ijt} = \frac{\sum_{p=1}^{t-1} b(t-p) X_{ijp}}{\sum_{p=1}^{t-1} b(t-p)}$$

where $b \in (0, 1]$ represents the decay factor, and $X_{ijp}$ represents the outcome of practice $p$ (1 if successful and 0 if failed) for student $i$ on skill $j$. In the original R-PFA paper, authors used three ghost/synthetic practices (Galyardt & Goldin, 2014).

Note that in the first R-PFA paper each item was linked to a single skill, losing one of the main original benefits of PFA. To the best of our knowledge, there is no previous work on extending R-PFA to create a variation which can handle data containing multi-skill items. However, recency has been considered, including adding time-based weights to components of existing models and incorporating a weighted proportion for failures (Pavlik et al., 2021).

Further work is seen in Logistic Knowledge Tracing (LKT) (Pavlik et al., 2021), a logistic regression-based framework that can enable multiple components from different existing models, including PFA (Pavlik et al., 2009), PFA-Decay (Gong et al., 2011), R-PFA (Galyardt & Goldin, 2014), and PPE (Walsh et al., 2018). LKT showcased a suite of components that could be combined to form new models, some of which incorporated the notion of recency and decay. Two of the comparison models we implement, Alg1 and Alg2 (described further in
Section 6), are based on components described in LKT. Important components from (Pavlik et al., 2021) that we used in our comparison models are as follows:

- **Intercept for each KC/skill**, which is a simple linear model intercept. Used in Alg1 and Alg2.
- (Alg1) Log performance (logsucc, logfail), which is the log-transformed performance factor (the total successes or failures), representing declining marginal returns, e.g. \( \ln(s_{ij}) + \ln(f_{ij}) \).
- (Alg2) Exponential decay of proportion, which uses the prior probability correct for each knowledge component, as in the R-PFA model. This uses a parameter to describe the exponential rate of decay, or recency, for observations of a knowledge component.
- (Alg2) The LKT version of Recency, defined as the power log decay applied to the time interval since the previous encounter with the KC. This feature considers only the just prior observation and simulates performance improvement when the prior practice was recent.

In this paper we propose a model called MemDec that captures memory decay and spacing, and compare its performance to PFA, R-PFA, and components of LKT, while accounting for multiple skills per item. We implement and analyze two methods of incorporating decay through practice order and through time windows. This proposed algorithm considers the effects of spacing between practices, thus modeling the spacing effect when predicting student performance. It also studies multiple ways of inducing decay, whereas previous studies focus only on one method, mostly practice-order. More study of the spacing effect in knowledge tracing is warranted. While a few algorithms have utilized the notion of time windows (Choffin et al., 2019; Lindsey et al., 2014), their time windows overlap and are defined in a relatively restrictive and limited manner. Additionally, (Choffin et al., 2019; Lindsey et al., 2014) do not consider time elapsed between practices when modeling spacing, whereas our proposed model MemDec Spacing considers the time elapsed between each current and previous practice, which we believe is crucial for modeling spacing accurately. Also, in MemDec, this is done for both practice-order and time-window variations. For MemDec and MemDec Spacing with time windows, we decay the weight of a practice based on the time window the practice falls into. Using equivalent, disjoint time windows allows for consistency in the exponential rate of decay through time. Additionally, we conduct an in-depth analysis on the impact of changing the decay factors and the number of ghost practices.

### 4. MemDec Algorithm

#### 4.1 MemDec (PFA Memory Decay)

We propose a new model, MemDec, a variation of the PFA algorithm inspired by R-PFA components, that can also be seen as fitting within the LKT framework. In MemDec, memory decay is applied to both successful and failed practices, and the model can be used with multi-skill items (whereas the original R-PFA only supports single-skill items).

In addition to this, we build on the approach in (Maier et al., 2021), which splits skills into "common" or "rare" categories, so that it can be applied for learning systems where some skills are rare. Rare skills can occur when -- for example -- items are tagged with skills that represent prerequisites not taught in the courseware. When training PFA with datasets containing rare skills, several challenges including degenerate parameters can occur (Maier et al., 2021). Depending on how rare some skills are, there might not be enough data points to precisely estimate parameters when training a model. (Maier et al., 2021) proposes a PFA variant that splits out common and rare skills. When training a model, each common skill has its own set of parameters, while all rare skills are combined into a single common set of default parameters, improving predictions and reducing model degeneracy (Maier et al., 2021).

The formula for MemDec is given below (note that as for PFA and R-PFA, \( m \) will be inputted into a logistic function to obtain a prediction, \( p(m) \)):

\[
m(i,j, KC, RS, RF) = \sum_{j \in \text{common } KC} \left( \beta_j + \gamma_j RS_{ij}t_{ij} + \delta_j RF_{ij}t_{ij} \right) + \sum_{j \in \text{rare } KC} \left( \beta_d + \gamma_d RS_{ij}t_{ij} + \delta_d RF_{ij}t_{ij} \right)
\]
\[
RS_{ijtij} = \frac{\sum_{p=0}^{t_{ij}-1} b_z (t_{ij}-p) X_{ijp}}{\sum_{p=0}^{t_{ij}-1} b_z (t_{ij}-p)} \\
RF_{ijtij} = \frac{\sum_{p=0}^{t_{ij}-1} b_f (t_{ij}-p) (1-X_{ijp})}{\sum_{p=0}^{t_{ij}-1} b_f (t_{ij}-p)}
\]

Where \( i \) represents a student, \( j \) represents a skill, \( t_{ij} \) is the current trial (i.e. practice) number student \( i \) is on with skill \( j \), \( X_{ijp} \) represents the correctness of the practice (i.e. it is 1 if student \( i \)'s practice \( p \) with skill \( j \) was successful, and 0 otherwise). Constant \( b_z \in (0,1] \) is the decay rate for successful practices, and constant \( b_f \in (0,1] \) represents the decay factor for failed practices. RS and RF represent the recency-weighted proportions of past successes and past failures, respectively. The values for \( \beta_j \), \( \gamma_j \) and \( \delta_j \) are parameters that are learned for each skill \( j \) during the training. Parameters \( \beta_d \), \( \gamma_d \) and \( \delta_d \) are the default parameters learned for the rare skills. Only one value is learned for each parameter for the set of rare skills.

Like R-PFA, MemDec can incorporate ghost/synthetic practices (imaginary practices that improve model performance initially) in the RS and RF formulas. To allow for ghost practices, we start \( p \) from a negative number (instead of starting from 0). In the original R-PFA formula, the authors proposed three ghost practices (all failed practices). As in LKT, we investigate using no ghost practices, two ghost practices (one successful and one failed), and three ghost practices (all failed). The main differences between R-PFA and the MemDec base variant is that MemDec does not use a total term; instead, it contains a component that takes into consideration the weighted proportion of failed practices, giving more weight to recent ones. Also, MemDec can handle multi-skill items whereas R-PFA is designed for only single-skill items. Additional, R-PFA only considered a practice order approach, whereas one MemDec variant considers a time window approach.

In MemDec, as in most models that incorporate the notion of decay, the order of practices plays an important role. Every time the student completes a new practice, the model introduces more decay to previous practices. This means that the more practices the student has, the more decay is applied to older practices. In this approach, while decay is incremented by more practice, the time elapsed between practices does not affect the calculation of decay.

Practices can be separated by different amounts of time, from seconds to months. It is unlikely that substantial forgetting will occur with small amounts of elapsed times, such as seconds or minutes. This could be a limitation to the practice-order approach, if the amount of elapsed time can vary considerably. Thus, we propose a variation of MemDec that uses a time window, a constant duration of time (for example: 1 day) in which items answered within the same time window are given equal decay. This accounts for memory decay not occurring instantaneously. Practices answered in time windows farther from the current practice can be expected to have decayed more than practices from more recent time windows. For this variant, MemDec's RS and RF formulas become:

\[
RS_{ijtij} = \frac{\sum_{p=0}^{t_{ij}-1} b_z \text{timewindow}(time(t_{ij})-time(p)) X_{ijp}}{\sum_{p=0}^{t_{ij}-1} b_z \text{timewindow}(time(t_{ij})-time(p))} \\
RF_{ijtij} = \frac{\sum_{p=0}^{t_{ij}-1} b_f \text{timewindow}(time(t_{ij})-time(p)) (1-X_{ijp})}{\sum_{p=0}^{t_{ij}-1} b_f \text{timewindow}(time(t_{ij})-time(p))}
\]

4.2 MemDec Spacing

The spacing effect has to do with the temporal distribution of practices linked to the same skill. If minimal time has elapsed between practices, the learning is said to be massed. Existing research suggests that if practices are spaced out, information is retained longer in memory (Pavlik & Anderson, 2005). In this section, we investigate an extension of MemDec that incorporates the notion of the spacing effect into knowledge estimation. The MemDec model was adjusted to use \( b_z \) values that are calculated based on a formula which takes into account how spaced apart practices are. The values of \( b_f \) are calculated in a similar manner:
\[ b_s(t_{ij}) = \begin{cases} 
    b_{s_{\text{min}}}, & \text{if no prev practices, or elapsed time} = 0 \\
    \min(b_{s_{\text{min}}} + \log_z(\text{time}(t_{ij}) - \text{time}(t_{ij} - 1)), b_{s_{\text{max}}}), & \text{otherwise} 
\end{cases} \]

\[ b_f(t_{ij}) = \begin{cases} 
    b_{f_{\text{min}}}, & \text{if no prev practices, or elapsed time} = 0 \\
    \min(b_{f_{\text{min}}} + \log_z(\text{time}(t_{ij}) - \text{time}(t_{ij} - 1)), b_{f_{\text{max}}}), & \text{otherwise} 
\end{cases} \]

where \( b_{s_{\text{min}}} \in (0,1] \) and \( b_{s_{\text{max}}} \in (0,1] \) represent the min and the max values that we allow for \( b_s \). Constants \( b_{f_{\text{min}}} \in (0,1] \) and \( b_{f_{\text{max}}} \in (0,1] \) represent the min and the max values that can be used for \( b_f \). Constant \( z \) is the base of the logarithm, and expression \( \text{time}(t_{ij}) - \text{time}(t_{ij} - 1) \) calculates the elapsed time between the current practice with skill \( j \) and the previous practice with skill \( j \) performed by student \( i \). Note that if the student has no previous practices with a skill, or the elapsed time is 0 (elapsed time of 0 can occur in real systems if timestamps are not captured at enough granularity), then we assign the minimum values. MemDec Spacing could be used with either practice order (p.o.) or with a time window.

5. Dataset

For the experiments presented in this article, we used data from Reveal Math Course 1, a McGraw-Hill digital math product that covers grade 6 US math curriculum. The items from the assessments from this data are tagged with one or more skills. The data we used came from two Midwestern school districts and one Southwestern school district. One of these school districts is within a large U.S. city where over half of students are classified as Black, around 10% of students are classified as Hispanic, and a fifth of families live under the poverty line. A second district is within a small town where around 5% of students are classified as Black, around 90% of students are classified as White, and around 10% of families live under the poverty line. A third is within a larger town where just over half of students are classified as Hispanic, just under half of students are classified as White, and about 3/4 of families live under the poverty line. All use the NGA Center/CCSSO Common Core Standards.

Extracted data spans between August 2019 and May 2021. There are 4,363 unique items; 2,009 are tagged with multiple skills. The items include multiple choice, fill in the blank, and entering equation items. Overall, the dataset had 71 unique skills which were linked to the items from the dataset. Out of these skills, 42 were classified as common (at least 200 students with at least 3 practices – Maier et al., 2021) and the remaining 29 were classified as rare.

The dataset has 489,359 datapoints. Datapoints represent students’ responses and their normalized scores (1 if the response is correct, 0 if incorrect). 1.25% of the datapoints contained a partially correct score, which were treated as 0 for the purposes of this analysis. For the experiments, we split our dataset into training and testing sets. We randomly selected about 20% of the students (647 students, 98,604 data points, 64 skills) for the testing set, leaving 80% of the students (2,588 students, 390,755 data points, 71 skills) for training.

6. Experimental Results

For validation, we ran several experiments using the proposed approaches from this article, as well as other existing algorithms. For comparison reasons, we implemented the original PFA (Pavlik et al., 2009) with adjustments to handle rare skills as described in (Maier et al., 2021). We call this the Baseline model. We also implemented other algorithms to benchmark against MemDec and MemDec Spacing: R-PFA (Galway & Goldin, 2014), and two algorithms that were inspired by models from LKT (Pavlik et al., 2021). We provide information on those in the Comparison Models sub-section. For all models, we allowed for multi-skill items by using a summation factor across multiple skills linked to an item.

In an effort to study the differences and the effectiveness between each model, we calculated the AUC and RMSE validation metrics. Also, we present validation results for different groups of datapoints within the testing dataset: “all data” means we validated against all datapoints; “1+ non-default” means that we only used datapoints for which the item was
linked to at least one common skill; “only non-default” means we only used datapoints whose items were linked with exclusively common skills; “1+ default” means we only used datapoints whose items were linked to at least one rare skill; and “only default skills” means datapoints with items solely tagged to rare skills.

6.1 Baseline Results

We trained a model that learned three parameters for each of the 42 common skills and three parameters for the rare skills. The validation results are presented in Table 1 below:

<table>
<thead>
<tr>
<th>Category</th>
<th># of Data Points</th>
<th>AUC, RMSE (Baseline)</th>
<th>AUC, RMSE (MemDec Practice Order)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>98604</td>
<td>0.6975, 0.4443</td>
<td>0.7679, 0.4076</td>
</tr>
<tr>
<td>1+ non-default</td>
<td>97460</td>
<td>0.6959, 0.4446</td>
<td>0.7675, 0.4074</td>
</tr>
<tr>
<td>Only non-default</td>
<td>95859</td>
<td>0.6952, 0.4444</td>
<td>0.7677, 0.407</td>
</tr>
<tr>
<td>1+ default</td>
<td>2745</td>
<td>0.7554, 0.4417</td>
<td>0.7628, 0.431</td>
</tr>
<tr>
<td>Only default</td>
<td>1144</td>
<td>0.8083, 0.4163</td>
<td>0.8077, 0.4244</td>
</tr>
</tbody>
</table>

6.2 MemDec Models Results

To study the difference between methods used to govern the increase in decay, we implemented and tested a variation of MemDec that used the practice-order approach, as well as a variation that used the time-window method. For the decay factors \( b_x \) and \( b_y \) we tried several combinations of values from \([0, 1] \). While other combinations gave similar results, the best were obtained with \( b_x = 0.6 \) and \( b_y = 0.7 \). We present this model’s results in Table 1. We observe a significant improvement when compared with the Baseline model. By incorporating the notion of memory decay, MemDec achieved an AUC of about 0.77 on all testing datapoints, whereas the baseline reached only an AUC of about 0.7. Significant improvements were observed in all other categories of datapoints, except for categories involving default skills, for which the two models achieved similar performance, likely because rare skills do not contain enough datapoints in the dataset for our model to substantially learn from.

We also ran experiments with a time-window of 14d (14 days), 7d, 2d, and 1d (see Table 2). The best results were observed for a 1d time window with an AUC equaling 0.756 across all datapoints from the testing set, which is slightly lower than the AUC of the practice-order model. The 2d window model obtained an AUC of 0.754 for all data, the 7d window an AUC of 0.749, and the 14d window model an AUC of 0.747. This demonstrated that for this dataset, the model that uses the practice-order approach performs slightly better. For the time-window variation we observed that the smaller the window, the better the results.

<table>
<thead>
<tr>
<th>Category</th>
<th>AUC, RMSE (1d)</th>
<th>AUC, RMSE (2d)</th>
<th>AUC, RMSE (14d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>0.7561, 0.4126</td>
<td>0.7541, 0.4134</td>
<td>0.7468, 0.4166</td>
</tr>
<tr>
<td>1+ non-default</td>
<td>0.7557, 0.4124</td>
<td>0.7536, 0.4132</td>
<td>0.7462, 0.4164</td>
</tr>
<tr>
<td>Only non-default</td>
<td>0.7557, 0.4119</td>
<td>0.7536, 0.4127</td>
<td>0.746, 0.416</td>
</tr>
<tr>
<td>1+ default</td>
<td>0.7564, 0.4346</td>
<td>0.7552, 0.4343</td>
<td>0.7556, 0.4357</td>
</tr>
<tr>
<td>Only default</td>
<td>0.8053, 0.4287</td>
<td>0.8018, 0.4294</td>
<td>0.7992, 0.4311</td>
</tr>
</tbody>
</table>

To study whether modeling a combination of both decay and spacing could further improve the predictions, we ran experiments with the MemDec Spacing model. We experimented with different values for the hyperparameters which represents the lower and upper bounds of the decay factor, and for practice order we obtained very similar results compared to the non-spacing practice order MemDec models. By applying practice order and
parameters \((b_{s_{\text{min}}}=0.55, b_{s_{\text{max}}}=0.65, b_{f_{\text{min}}}=0.7, b_{f_{\text{max}}}=0.7, \log z = \log_2 )\), shown in Table 3 left column we obtained a slightly better result than MemDec without spacing, with an overall AUC of 0.768. For this approach and time window of 1d (1 day) the AUC was 0.756, which is slightly lower than the AUC when using practice-order (p.o.). Large time windows performed more poorly still. Overall, with time windows, MemDec Spacing gave slightly poorer results than the MemDec model. This finding may be due to certain properties of the dataset we use. Many datapoints are not spaced apart more than a few seconds in time, which would cause incorporating the effects of spacing (through time windows) to have negligible effects on the model’s calculation of student knowledge. When using practice order to represent decay, the effect of spacing seems to be negligible. It is possible that if a different dataset with more widely spaced practices is used, the effect of spacing on MemDec with practice order might be more beneficial. In Section 7 we will discuss the interpretation of these results further.

<table>
<thead>
<tr>
<th>Category</th>
<th>AUC, RMSE (p.o.)</th>
<th>AUC, RMSE (1d)</th>
<th>AUC, RMSE (7d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>0.768, 0.4076</td>
<td>0.7558, 0.4127</td>
<td>0.7492, 0.4156</td>
</tr>
<tr>
<td>1+ non-default</td>
<td>0.7675, 0.4074</td>
<td>0.7554, 0.4125</td>
<td>0.7487, 0.4155</td>
</tr>
<tr>
<td>Only non-default</td>
<td>0.7678, 0.4077</td>
<td>0.7554, 0.4121</td>
<td>0.7485, 0.4151</td>
</tr>
<tr>
<td>1+ default</td>
<td>0.7635, 0.4307</td>
<td>0.7569, 0.4347</td>
<td>0.7559, 0.4351</td>
</tr>
<tr>
<td>Only default</td>
<td>0.8087, 0.4245</td>
<td>0.8038, 0.4291</td>
<td>0.7992, 0.431</td>
</tr>
</tbody>
</table>

6.3 Comparison Models

We compare our models with R-PFA (Galjardt & Goldin, 2014), as well as two algorithms that incorporate components from LKT (Pavlik et al., 2021), which we call Alg1 and Alg2. The \(m\) function of each model is input into the sigmoid function, to get a probability value between 0 and 1. Alg1 contains a recency component that captures the elapsed time \((t)\) between current and previous practice of the student \(i\) with skill \(j\) raised to a decay factor \(d\). It also takes the natural logarithm of the number of successes \(s_{ij}\) and number of failures \(f_{ij}\).

Alg 2 contains a component that uses a weighted proportion of previous practices along with a parameter \(b\) that represents the exponential rate of decay, and ghost parameters.

\[
\text{Alg1: } m(t; j \in KC; s; f) = \sum_{j \in KC} (\beta_j + \gamma_j \ln(s_{ij}) + \rho_j \ln(f_{ij}) + \alpha_j t_{ij}^d)
\]

\[
\text{Alg2: } m(t; j \in KC; s; f) = \sum_{j \in KC} \left( \beta_j + \gamma_j \frac{\sum_{p=1}^{t-1} b^{(t-p)} x_{ij}}{\sum_{p=1}^{t-1} b^{(t-p)}} \right)
\]

Because many of these models used two (1 failed, 1 successful) ghost practices, or
three (3 failed) ghost practices, we also implemented and ran experiments with MemDec and MemDec Spacing using this combination of two or three ghost practices. For MemDec, the presence of ghost practices had a negligible influence on results. The results of R-PFA with different numbers of ghosts (successes and failures) are given in Table 4. These findings show that MemDec significantly outperformed R-PFA, regardless of the number of ghost practices. The number of ghost practices did not seem to have a major influence on the R-PFA results. Alg1 performed better than R-PFA, with an AUC of 0.7234 for all test data points. Alg2, with both two and three ghost practices, performed better than Alg1, with an overall AUC of 0.747 for two ghost practices and 0.739 for the model with three ghost practices. These results, shown in Table 5, are still worse than MemDec and MemDec Spacing.

<table>
<thead>
<tr>
<th>Category</th>
<th>3 ghost (0s, 3f)</th>
<th>2 ghost (1s, 1f)</th>
<th>0 ghost</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>0.6065, 0.4639</td>
<td>0.6067, 0.4638</td>
<td>0.6067, 0.4638</td>
</tr>
<tr>
<td>1+ non-default</td>
<td>0.6051, 0.4637</td>
<td>0.6053, 0.4636</td>
<td>0.6053, 0.4636</td>
</tr>
<tr>
<td>Only non-default</td>
<td>0.6069, 0.463</td>
<td>0.6071, 0.463</td>
<td>0.6071, 0.463</td>
</tr>
<tr>
<td>1+ default</td>
<td>0.574, 0.492</td>
<td>0.574, 0.492</td>
<td>0.5743, 0.4919</td>
</tr>
<tr>
<td>Only default</td>
<td>0.7614, 0.4802</td>
<td>0.7614, 0.4802</td>
<td>0.7615, 0.4801</td>
</tr>
</tbody>
</table>
Table 5. Alg1 and Alg2, AUC and RMSE

<table>
<thead>
<tr>
<th>Category</th>
<th>Alg1, 0 ghost</th>
<th>Alg2 (1s, 1f)</th>
<th>0 ghost</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>0.7235, 0.4258</td>
<td>0.7471, 0.4208</td>
<td>0.7395, 0.4227</td>
</tr>
<tr>
<td>1+ non-default</td>
<td>0.7223, 0.4258</td>
<td>0.7469, 0.4205</td>
<td>0.7387, 0.4226</td>
</tr>
<tr>
<td>Only non-default</td>
<td>0.7224, 0.4253</td>
<td>0.7476, 0.4199</td>
<td>0.7395, 0.4222</td>
</tr>
<tr>
<td>1+ default</td>
<td>0.7492, 0.4438</td>
<td>0.735, 0.4509</td>
<td>0.7279, 0.4485</td>
</tr>
<tr>
<td>Only default</td>
<td>0.8234, 0.428</td>
<td>0.8294, 0.4489</td>
<td>0.8191, 0.4355</td>
</tr>
</tbody>
</table>

Overall, MemDec and MemDec Spacing outperformed all other models implemented in this study, including PFA, R-PFA, Alg1, and Alg2. We find that the practice-order variation of MemDec Spacing and MemDec provided the best predictions, with a minimal higher performance seen in MemDec Spacing. Both were followed by the time-window MemDec variation with a slightly more significant difference. While MemDec Spacing with time-window was outperformed by MemDec with time-window, it was still more effective than any other tested models in this experiment. The practice-order model was able to estimate student knowledge much more accurately than Baseline PFA or R-PFA. Additionally, within the MemDec variants, practice-order models were more effective than time-window models, and ghost practices had a negligible effect on performance predictions.

7. Discussion and Conclusions

In this work we studied the cognitive science concepts of memory decay and the spacing effect in the context of variants of the Logistic Knowledge Tracing framework. We created a new algorithm called MemDec which expands on R-PFA, a variation of PFA that incorporates decay. Despite the early emphasis on multi-skill items being a strength of PFA (Pavlik et al., 2009), to the best of our knowledge, there is no previous work on R-PFA or other time-involved extensions that looked at data containing multi-skill items, although components in LKT can handle multi-skill items (Pavlik et al., 2021). Addressing this limitation increases the algorithm’s relevance to real-world educational systems where items are associated with multiple skills. We further expanded MemDec to capture the spacing effect in our model MemDec Spacing.

We also studied different ways of modeling decay, through the order of practices and by intervals of time elapsed between practices (time windows). To the best of our knowledge, previous extensions of PFA and LKT components mostly focused on a practice-order approach. We tested whether different values of the decay factor led to improved model predictions, in all variations. Our new algorithms were compared against two comparison algorithms based on LKT components, Alg1 and Alg2, as well as against PFA and R-PFA.

The results of this study showed that MemDec and MemDec Spacing outperformed all other comparison models. Practice-order MemDec variations showed better results than time-window variations. We investigated different time window sizes, from 1d to 14d windows, and found that smaller time-windows achieved better results. The study also found that modeling decay with the spacing effect did not seem to provide an advantage over solely modeling decay. For practice-order, MemDec Spacing performed slightly better than MemDec. Across time windows, MemDec outperformed MemDec Spacing by a small amount.

However, these findings may be due to the relatively massed nature of the current dataset. Therefore, it may be valuable for future work to compare these models within datasets containing more spaced items, to determine whether the time-window approach could be beneficial over practice-order in this situation. This would also show whether incorporating the spacing effect along with decay can have a bigger impact on prediction for such datasets.

Another area of future work involves looking into how well the approaches presented perform at predicting retention long-term. Finally, future work in this area may benefit from going beyond simply assessing predictive goodness to assessing the practical implications of when instructors are told a student has mastered a skill, when in fact they have forgotten it.

Overall, the fact that MemDec and MemDec Spacing outperformed the other models highlights the importance of capturing cognitive science principles such as memory decay and spacing when modeling student knowledge and predicting future performance. The analysis
conducted also showcases the difference in model performance between increasing decay by either order or through time. The results show that the proposed models are suitable knowledge tracing approaches for real-world adaptive learning systems with multi-skill items, where the real possibility of students forgetting skills can significantly impact the results.

References


Developing a Video-based e-Learning System Incorporating a Fill-in-the-blank Question-type Concept Map

Takumi HASEGAWAa & Tessai HAYAMAa*

a Nagaoka University of Technology, JAPAN
*t-hayama@kjs.nagaokaut.ac.jp

Abstract: E-learning, which can be used anywhere and at any time, is very convenient and has been introduced to improve learning efficiency. However, securing a completion rate has been a major challenge. Recently, the learning forms of e-learning require learners to be introspective, deliberate, and logical and have proven to be incompatible with many learners with low completion rates. In this study, we propose an e-learning system that incorporates a fill-in-the-blank question-type concept map to deepen learners' understanding of learning contents while watching learning videos. The proposed system promotes active learning reflectively and logically by allowing learners to answer blank question labels on concept maps from video content and labels associated with the question labels. We conducted an experiment with 14 subjects to evaluate our proposed system by comparing it with a conventional video-based e-learning system and confirmed that the proposed system promotes active learning while watching lecture videos and is effective in improving the comprehension of learning content for learners who are unfamiliar with video-based learning.

Keywords: e-Learning Support, Visual Thinking Tool, Video-based Learning

1. Introduction

In online video-based learning, a learner typically does not have to be in direct contact with the class at the same time, which has significant advantages. It supports learning anywhere and anytime via a mobile phone or a computer and enables access to learning that has never been available before. Recently, although online learning courses have been increasingly marketed to adults and students, the high dropout rate from such courses remains a concern to educational institutions. Much research has shown a high dropout percentage rate of learners participating in online courses compared with learners in a face-to-face classroom; e.g. more than 95% of learners dropped out from MOOC (Alraimi, Zo, & Ciganek (2015)) and approximately 78% of learners dropped out from Open University (Simpson (2013)).

According to a survey report (Tominaga & Koga (2014)), the reasons for learners dropping out of online learning courses are classified into two major problem categories: learners' procrastination and incompatible learning style to e-learning. The first problem category entails that some learners cannot effectively manage the time and learning progress of e-learning by themselves, which makes it difficult to complete e-learning courses. To solve the first problem category, some e-learning institutions have introduced online tutoring and mentoring to provide individual support to such learners (Huang, Lin, & Huang (2012)). Others have identified learners who are likely to drop out by analyzing their learning activities from system logs (Mduma, Kalegele, & Machuve (2019)). These approaches are effective ways to maintain the retention rate of e-learning courses (Grau, & Minguillon (2013)). Meanwhile, the second problem category is that multimedia content provided by e-learning systems generates cognitive load effects on learners compared with typical face-to-face learning in the classroom (DeLeeuw & Mayer (2008)). The cognitive load effects in e-learning reduce the learning performance of learners, increasing the tendency of learners to drop out. To solve the second problem category, some researchers have designed e-learning systems with functions to match the learning style of e-learning (Kizilcec, Papadopulolos, & Sritanyaratana (2014)).
Mayer(2005); Guo, Philip, Kim, & Rubin (2014)). Despite creating a better approach than the normal e-learning style, many recent e-learning systems still adopt a simple form with a video in the center of the interface. Moreover, some researchers have investigated the learning style of e-learning and demonstrated that both reflective and active learning styles are relevant to enhance the e-learning progress and that sensory learners are probably adaptive to the learning style of e-learning(Huang, Lin, & Huang(2012); Dalmolin, Mackeivicz, Pocharpski, Pilatti, & Santos (2018)). Therefore, it is required for a learner to perform learning with metacognitive skills and a step-by-step thinking method during e-learning. However, it is difficult for many learners to use their skills because current video-based learning affords passive learning, and it is challenging for learners to leave the concepts of learning content while watching the video. Furthermore, learners who have dropout tendencies cannot effectively understand the learning content from the beginning of an e-learning course (Simpson (2013)). If the e-learning system has support to match a learner’s thinking method in e-learning and learning style, which e-learning requires, the dropout rates in each e-learning course will be reduced.

Therefore, in this study, we propose an online video learning interface that provides a fill-in-the-blank question-type concept map. A concept map is an effective way to provide complex concepts with graph representations and frequently used in educational institutions (Novak (1990)). However, it is almost impossible for them to create an appropriate concept map while watching a lecture video(Hayama & Sato(2020)). Thus, we employed a fill-in-the-blank question-type concept map in which some labels that are important for understanding learning contents are punched out from a completed concept map, encouraging learners to think about them while watching the video. To achieve this, we also developed a method for extracting labels for fill-in-the-blank questions from the concept map. We conducted an experiment with 14 subjects to evaluate our proposed system by comparing it with a conventional system with a concept map and confirmed that the proposed system promotes active learning while watching a lecture video and is effective in improving the comprehension of learning contents for learners who are unfamiliar with video-based learning.

2. Related Studies

A concept map is a useful tool for supporting learners’ deep understanding, evaluating their understanding, and so on (Novak (1990); Schmid & Telaro (2015)) as it can visually represent the relationships between various concepts. Some researchers have used concept maps to support learning by reflecting on the learning content of a class, complementing or deepening the understanding of the learning content after the class. For example, Cimolino & Kay (2002) confirmed the effectiveness of concept maps for teachers to assess students’ understanding of concepts in a subdivided learning area by having students create and analyze concept maps to recognize conceptual understandings and misunderstandings of the learning content. As it is difficult for all learners to create proper concept maps, some researchers have developed a method using Kit-Build concept maps, where experts prepare components with labels and links in advance, and learners assemble them. For example, Hirashima, Yamasaki, Fukuda, & Funaoi (2011) demonstrated that Kit-Built concept maps can be used to diagnose a learner’s comprehension of learning content and to compare the concept maps of learners with those of experts, respectively, for learning assessment. Hayashi (2019) presented more examples in which a Kit-Built concept map was highly useful in actual classroom teaching and cooperative learning. Conversely, some researchers have developed systems incorporating fill-in-the-blank concept maps, where some labels are masked from a concept map created by an expert. Chang, Sung, & Chen (2001) developed a system using fill-in-the-blank concept maps. They confirmed that the system was more effective than conventional concept maps in terms of students’ understanding of learning content through the task of creating a concept map of learning content after the class.

Nevertheless, with the recent spread of online learning, technologies using concept maps while watching lecture videos have been developed. For example, Liu, Kim, & Wang (2018) developed a system incorporating concept maps as navigation while watching lecture videos. In their study, they proposed a concept map creation method by aggregating multiple
concept maps created by general users and confirmed that the concept maps created by their method contained contents similar to those created by experts. However, their method only allows learners to pore over concept maps while watching a lecture video. To promote a more active learning style, learners need to interact when creating concept maps while watching lecture videos. In this study, we propose a learning support system incorporating a concept map with fill-in-the-blank questions, which makes a learner answer questions about important points of the learning contents while watching a lecture video. To achieve this, we develop a method for identifying the fill-in-blank labels on the concept map to understand the learning contents. To the best of our knowledge, this is the first study on the above approach.

3. Developed System

We developed a video-based e-learning system using a concept map with fill-in-the-blank questions. As depicted in Figure 1, the developed system consists of a lecture video-watching area, a concept map display area, and an answer input area for fill-in-the-blank questions.

In the video-watching area, lecture videos, including lecture slides and lecture speeches, are provided. These lecture videos can be played and stopped using buttons, and the playback position can be specified using the playback bar. In the concept map display area, the locations of multiple labels on the concept map created by the lecturer or other experts are automatically punched out and displayed. As shown in the concept map display area, the yellow blank labels have been automatically punched out of the concept map. Each label is automatically associated with the playback position of the lecture video that may contain the content of the label. To make the user notice the label(s) associated with the playback position of the lecture video, the label(s) is/are colored orange, and a link to the corresponding video playback position is also set.

In the answer input area for the fill-in-the-blank questions, a label candidate list is provided, as well as an execute button for matching the answers. In the label candidate list, a list of punched-out labels is displayed. When a user presses the execute button to match the answers, the system checks whether the item inserted into the blank label in the concept map is correct or not, and if the wrong item is inserted into the blank label, the label turns red to notify the user that it is a misunderstanding.

While watching a lecture video, the user checks the part of the concept map related to the learning contents being watched, considers the learning contents corresponding to the blank label, and inserts the item into the label by selecting it from the candidate label list. The
user then confirms whether his/her answers are correct, and if there is/are wrong answer(s), the user reconfirms the corresponding learning contents in the video. Thus, the system enables the user to always be aware of important points of lecture contents to pay attention to and to watch the lecture video while confirming information related to the important points.

**Label extraction for fill-in-the-blank questions on concept maps:**

We created heuristic rules to automatically extract labels for fill-in-the-blank questions from a concept map. To extract fill-in-the-blank question labels, labels related to important parts of learning content are identified on the basis of the number of links to each label and the kinds of link labels, and blank labels should be recognized from the surrounding labels. As link labels for concept maps, simplified by trial and error, the labels of “Method,” “Description,” “Function,” “Example,” “Contents,” “Advantage/Disadvantage,” and other user’s definition phrases were used. The label extraction rules for fill-in-the-blank questions are listed below.

(a) If a label outputs a link with a `Description` label, the label is left blank because the answer to the label can be inferred from the output link label that describes it.

(b) If a label outputs a link with a label that appears only once in the concept map, the linked label is left blank because the content of the linked label is inferred from contents associated with the unique user’s defined link label.

(c) If a label outputs multiple labels linked with a link label of `Method`, `Description`, `Function`, or `Contents`, the odd-numbered label(s) from the left side of the linked labels is/are left blank because the characteristics of the contents of the linking label are inferred from the linked labels. However, if a blank label has no output link, the blank to the label is canceled.

(d) If a label outputs link(s) with a `Function` label, the linked label(s), which have more than one output link(s), is/are left blank because the content of each label can be inferred from the contents of labels linking/linked to the label.

(e) If a label does not output a link and is linked from multiple labels, the label is left blank because the content of the label is inferred from the contents of multiple labels linked to the label.

(f) If a label outputs links with `Advantage` and `Disadvantage` labels, the labels to which each link of their labels is output are left blank because the contents of the label can be inferred from the link labels with its advantage and disadvantage.

(g) If a label outputs only a link(s) with an `Example` label(s), the label is left blank because the content of the label is inferred from the contents of the example(s) of the label.

4. Evaluation

4.1 Settings

We conducted an experiment to investigate the effectiveness of our proposed system in terms of comprehension of the learning content and interaction with the system. In this regard, the developed system and a video-based system that incorporates a concept map as it is were compared based on the comprehension test results and the operation history of the system.

Two lecture videos (approximately 50 min each) on the subject “Information Networks” were selected for the experiment. Fourteen undergraduate and graduate students majoring in information engineering, but who had never taken the subject, were recruited as subjects. For cross-validation, the subjects were divided into two groups (8 and 6 subjects, due to the absence of two subjects). The time limit for each lecture was set to 1 h, and each subject completed the lectures to his or her satisfaction. A comprehension test was then conducted for each lecture, consisting of 29 and 26 correct/incorrect questions on the learning contents of each lecture. Given the difference in the number of questions and the difficulty level of the lectures, the subjects’ deviation scores for each lecture were compared between the two systems for learning comprehension. The system operation history for each subject during the experiment was recorded with timestamps in a file, including the playback positions of the learning videos and the answers to the fill-in-blank questions on the concept maps. The
concept maps for Lectures 1 and 2 used in this experiment were similarly composed of 77 and 67 labels, 35 and 29 blanked labels, and 83 and 76 links, respectively. The concept maps and comprehension tests were created by the lecturer.

4.2 Results

The results of the comprehension test are shown in Figures 2 and 3, respectively.

![Figure 2](image2.png) **Figure 2.** Average deviation values of learners’ comprehension scores using the developed system/comparison system.

![Figure 3](image3.png) **Figure 3.** Deviation values of comprehension scores for each subject with lower performance when using the comparison system.

We calculated the deviation values of the subjects’ comprehension scores for each lecture and compared them separately for the proposed and comparison systems. As shown in Figure 2, there was little difference between the proposed and comparison systems. Conversely, the deviation values of the comprehension scores were examined for the seven subjects with lower performance when using the comparison system, which requires a conventional learning style for a learner. As shown in Figure 3, the results showed that the five subjects improved their performance when using the proposed system, with subject (C) exhibiting a significant difference in performance when using both systems and subject (E) exhibiting lower performance; thus, this indicates that subjects with low performance in the conventional system improved their understanding of the learning contents in video-based learning by using the proposed system.

Next, we confirmed that all of the system functions were used while watching the learning videos. All subjects provided correct answers to the fill-in-blank questions on the labels for each lecture. The number of trials for matching the correct answers to the fill-in-the-blank questions was 4.3 and 4.8 for the 77 and 67 questions in Lectures 1 and 2, respectively, indicating that the subjects answered some of the questions before jointly checking their correctness. The number of times the subjects intentionally moved the playback position of the lecture videos during learning was 29.4 and 12.4 for the proposed and comparison systems, respectively, indicating that the subjects moved the playback position of the lecture videos more frequently with the proposed system than with the comparison system, with a statistically significant difference.

The post-questionnaire on the proposed system showed good results, with all items being rated at 3.4 or higher out of 5. For the learning environment of the proposed system, “Satisfaction with his/her learning,” “Level of understanding of the learning contents,” “Sufficiency of the learning time,” and “Usability of the proposed system” were highly rated, indicating that the proposed system is a favorable learning environment to introduce. For the functions of the proposed system, “Effectiveness of the system function ‘concept maps with fill-in-the-blank questions’ in understanding the learning contents” was rated very high at 4.36 of 5. From the free description items, it can be inferred that the proposed system made it easier for the learners to watch and learn lecture videos by making them aware of what they need to pay attention to while watching the videos. In addition, we confirmed that the function of “the label extraction for the fill-in-the-blank questions on the concept map” was properly applied to
the concept maps of Lectures 1 and 2 because the labels for the fill-in-the-blank questions were related to the questions of the comprehension tests.

Overall, the proposed system promotes active learning by having learners answer fill-in-the-blank questions on a concept map while watching a lecture video and can provide learning effects to improve the understanding of learning contents for learners who do not fit the learning style of conventional video-based learning.

Acknowledgements
This work was supported by JSPS KAKENHI Grant Numbers 19K12264 and 23K11364.

References
Exploring Students' Adoption of ChatGPT as a Mentor for Undergraduate Computing Projects: PLS-SEM Analysis

Swapna GOTTIPATI*, Kyong Jin SHIM, Venky SHANKARARAMAN
School of Computing and Information Systems, Singapore Management University, Singapore
Swapnag,kjshim,venks@smu.edu.sg

Abstract: As computing projects increasingly become a core component of undergraduate courses, effective mentorship is crucial for supporting students' learning and development. Our study examines the adoption of ChatGPT as a mentor for undergraduate computing projects. It explores the impact of ChatGPT mentorship, specifically, skills development, and mentor responsiveness, i.e., ChatGPT's responsiveness to students' needs and requests. We utilize PLS-SEM to investigate the interrelationships between different factors and develop a model that captures their contribution to the effectiveness of ChatGPT as a mentor. The findings suggest that mentor responsiveness and technical/design support are key factors for the adoption of AI tools like ChatGPT. The study provides practical implications for educators seeking to incorporate AI as a mentor to support students doing computing projects and contributes to the broader understanding of the use of AI in education.

Keywords: Artificial Intelligence, ChatGPT, mentorship, computing projects, PLS-SEM analysis

1. Introduction

Recently, chatbots and virtual assistants gained popularity in education, providing personalized support to students (Mhlanga, 2023). ChatGPT, an AI language model by OpenAI, has been used in education (Sok et al., 2023) but not extensively as a mentor for computing projects. This study investigates the impact of ChatGPT mentorship, skills development, and responsiveness on its adoption as a mentor for undergraduate computing projects. Mentorship is crucial in student learning, especially in computing (Kuhail et al., 2023; Gottipati et al., 2023). However, traditional human mentors face limitations in availability and expertise (Eller et al., 2014). Chatbots like ChatGPT offer on-demand support, trained on vast data for tailored responses, making them attractive mentors. Still, their effectiveness in computing projects is unclear. Previous research focused on chatbots in education but not as mentors in computing (Kuhail et al., 2023). Variables like ChatGPT mentorship, skills development, and responsiveness in the context of computing projects need further exploration.

This paper examines the relationship between ChatGPT mentorship and its adoption in computing projects. Using the Partial Least Squares Structural Equation Modeling (PLS-SEM) statistical analysis method, we investigate the connections between various factors related to ChatGPT’s use as a mentor (Hair et al., 2021). It explores how factors like ChatGPT's effectiveness as a mentor, its usage in mentoring, and student skills development are interconnected. Ultimately, a comprehensive model is developed to understand how these factors contribute to ChatGPT's effectiveness as a mentor in computing projects. This study's implications are significant. It highlights chatbots' potential as mentors for computing projects, guides the design of effective mentorship programs, and contributes to the understanding of AI in education's benefits and limitations.
2. Related Work

Artificial Intelligence in Education (AIED) can enrich teaching and learning in higher education by personalizing student learning, automating assessment tasks, and even providing around-the-clock access to learning resources. In education, chatbots can be integrated into e-learning platforms that use Natural Language Processing (NLP) to interact with students by interpreting their questions and retrieving relevant information from the knowledge base module to assist students with revision (Clarizia et al., 2018). Interfacing with such conversational chatbots can provide students with the benefits of one-on-one mentoring. Active and frequent interaction between the student and the instructor impacts the student’s academic performance (Kuhail et al., 2023). Using chatbot software applications, institutions can scale one-on-one mentoring so that students have continuous support and can receive feedback 24/7. Benefits of chatbots for undergraduate students’ academic performance are demonstrated by Essel et al. (2022). In our work, we adopt the approaches from recent studies and apply them to analyze the use of ChatGPT to mentor students while doing their computing projects.

3. Hypothesis Development

In this section, we present the hypothesis of the study by identifying the factors of mentorship.

Technical mentorship: ChatGPT provides technical guidance to the student, covering data engineering, analysis, software architecture, coding standards, and best practices in software development (Castro et al., 2018). We assess the quality of advice provided by ChatGPT for completing technical tasks as a measure of technical mentorship.

Design mentorship: This construct involves ChatGPT’s guidance in design-related areas (Castro et al., 2018). It includes advising on software modules, analytical techniques, data sources, and data processing methods to achieve project objectives (Iacob & Faily, 2020). We evaluate ChatGPT’s effectiveness in helping students complete design-related tasks as a measure of design mentorship.

Mentor responsiveness: This construct focuses on ChatGPT’s prompt and effective response to the student’s needs and requests. Effective communication and understanding are crucial for successful problem-solving (Yin & Qiu, 2021). To assess mentor responsiveness, we measure the speed and quality of ChatGPT’s responses to the student’s questions.

Skills development: This construct focuses on the specific skills and knowledge gained by the student through ChatGPT’s mentorship. A mentor enhances the student’s professional and personal development by transferring technical and soft skills. We assess skills development by measuring the student’s perception of acquired skills after the mentorship.

Future adoption: This construct relates to students’ likelihood of using AI tools in future computing projects. Positive experiences increase the likelihood of future use, while challenges may discourage it. To assess future adoption, we measure student satisfaction and their intention to use ChatGPT in subsequent tasks.

We formulate below statistical hypothesis and test it with SMART-PLS 4 (Khoi & Van, 2018).

H1: ChatGPT technical mentorship has a positive impact on student technical skills development
H2: ChatGPT design mentorship has a positive impact on student soft skills development
H3: ChatGPT responsiveness has a positive impact on the future adoption of ChatGPT as mentor
H4: Technical skills development has a positive impact on the future adoption of ChatGPT as mentor
H5: Soft skills development has a positive impact on the future adoption of ChatGPT as mentor
4. Methodology

4.1 Participants

Participants in this study were undergraduate students from a "Text Mining and Language Processing" course at a university. The course lasted for 14 weeks, with a 12-week project component starting from week 3 and ending in week 14. The students worked in teams, applying skills in designing a text analytics application using Python and Java libraries. They were also required to demonstrate test cases during project presentations and complete a technical report. A total of 64 students were recruited for the survey questionnaire, with 58 fully completed surveys used for the PLS-SEM study.

4.2 Data Collection and Analysis

The survey instrument comprises several sections, each targeting different aspects of ChatGPT's mentorship. The technical mentorship section focuses on code-related solutions, while the design mentorship section examines project design approaches. The responsiveness section gauges student satisfaction with ChatGPT's guidance, and the skills development section assesses its impact on both technical and soft skills. Additionally, the survey explores students' future intentions regarding ChatGPT's use as a mentor. Finally, an open-ended question aids the qualitative analysis of the research question. We use the statistical software package SmartPLS 4 (Khoi & Van, 2018) for data analysis. The significance of relationships among variables is tested using regression coefficients, T-statistics, and P-values. The reliability and validity of the scale are tested by Cronbach's Alpha, Average Variance Extracted (AVE), and Composite Reliability (rho_c). A Cronbach's alpha coefficient greater than 0.6 ensures scale reliability (Hajjar, 2018). Rho_c should be better than 0.6 and AVE must be greater than 0.5 (Hair et al., 2021).

5. PLS-SEM Model Results and Analysis

5.1 Model Construction

Table 1 presents the variables used in the study and their descriptions.

<table>
<thead>
<tr>
<th>Construct Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUSE_1</td>
<td>ChatGPT provided code without logical errors for my project.</td>
</tr>
<tr>
<td>TUSE_2</td>
<td>ChatGPT provided code without missing steps for my project.</td>
</tr>
<tr>
<td>TUSE_3</td>
<td>ChatGPT provided clear solutions to my questions.</td>
</tr>
<tr>
<td>DUSE_1</td>
<td>ChatGPT supported me to find additional resources to support my project.</td>
</tr>
<tr>
<td>DUSE_2</td>
<td>ChatGPT supported me to generate new ideas or approaches for my project.</td>
</tr>
<tr>
<td>MSUP_1</td>
<td>ChatGPT responded to my questions and concerns during the project.</td>
</tr>
<tr>
<td>MSUP_2</td>
<td>ChatGPT provided sufficient support and guidance for my project.</td>
</tr>
<tr>
<td>MSUP_3</td>
<td>The feedback and suggestions provided by ChatGPT were helpful for my project.</td>
</tr>
<tr>
<td>TSKILL_1</td>
<td>ChatGPT increased my confidence in coding.</td>
</tr>
<tr>
<td>TSKILL_2</td>
<td>ChatGPT increased my confidence in technical report writing skills.</td>
</tr>
<tr>
<td>TSKILL_3</td>
<td>ChatGPT increased my levels of problem-solving skills.</td>
</tr>
<tr>
<td>TSKILL_4</td>
<td>ChatGPT made me less dependent on faculty/peers for coding help.</td>
</tr>
<tr>
<td>SSKILL_1</td>
<td>ChatGPT increased my levels of participation in team discussions.</td>
</tr>
<tr>
<td>SSKILL_2</td>
<td>ChatGPT increased my levels of quality in work submission.</td>
</tr>
<tr>
<td>SSKILL_3</td>
<td>ChatGPT increased my levels of time management in work submission.</td>
</tr>
<tr>
<td>FUT_1</td>
<td>ChatGPT is helpful as a mentor for my computing project.</td>
</tr>
<tr>
<td>FUT_2</td>
<td>I would use ChatGPT as a mentor for a future computing project.</td>
</tr>
</tbody>
</table>
5.2 Model Validity Checks, Results, and Analysis

5.2.1 Convergent validity

Table 2 shows that composite reliability varies from 0.762 to 0.894, which is above the preferred value of 0.5, demonstrating that the model is internally consistent. From Table 2, it can be observed that all the factors are reliable (Cronbach’s alpha > 0.64, AVE > 0.5) and have composite reliability (rho_c, rho_a) greater than 0.65. Therefore, all the factors are supported, and we can proceed with the model analysis in the next steps.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cronbach’s alpha</th>
<th>Composite reliability</th>
<th>Composite reliability</th>
<th>AVE</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUSE</td>
<td>0.674</td>
<td>0.869</td>
<td>0.848</td>
<td>0.738</td>
<td>Supported</td>
</tr>
<tr>
<td>FUT</td>
<td>0.644</td>
<td>0.67</td>
<td>0.847</td>
<td>0.735</td>
<td>Supported</td>
</tr>
<tr>
<td>MFB</td>
<td>0.69</td>
<td>0.716</td>
<td>0.829</td>
<td>0.62</td>
<td>Supported</td>
</tr>
<tr>
<td>SSKILL</td>
<td>0.666</td>
<td>0.96</td>
<td>0.84</td>
<td>0.728</td>
<td>Supported</td>
</tr>
<tr>
<td>TSKILL</td>
<td>0.707</td>
<td>0.727</td>
<td>0.812</td>
<td>0.526</td>
<td>Supported</td>
</tr>
<tr>
<td>TUSE</td>
<td>0.822</td>
<td>0.825</td>
<td>0.894</td>
<td>0.738</td>
<td>Supported</td>
</tr>
</tbody>
</table>

According to Chin, 2009, the variables that have path coefficients less than or equal to 0.5 should be excluded. Therefore, we removed SSKill_3 (path coefficient of 0.410) from the model. The final measurement model is depicted in Figure 1 (convergent validity is confirmed).

5.2.2 Discriminant Validity

The AVE can be used to verify the discriminant validity (Fornell & Larcker, 1981). If the squared correlation coefficients among the latent variables are less than the AVE values of the respective latent variables, then the discriminant validity is confirmed. Table 3 presents the data, and we observe that the correlation coefficients are less than the AVE indicating that variables are distinct and unrelated to each other, providing evidence for discriminant validity.
Table 3. Correlation coefficients among the variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>DUSE</th>
<th>FUT</th>
<th>MFB</th>
<th>SSKILL</th>
<th>TSKILL</th>
<th>TUSE</th>
<th>AVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUSE</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.738</td>
</tr>
<tr>
<td>FUT</td>
<td>0.576</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.735</td>
</tr>
<tr>
<td>MFB</td>
<td>0.536</td>
<td>0.605</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td>SSKILL</td>
<td>0.357</td>
<td>0.239</td>
<td>0.091</td>
<td>1</td>
<td></td>
<td></td>
<td>0.728</td>
</tr>
<tr>
<td>TSKILL</td>
<td>0.133</td>
<td>0.162</td>
<td>-0.01</td>
<td>0.558</td>
<td>1</td>
<td></td>
<td>0.526</td>
</tr>
<tr>
<td>TUSE</td>
<td>0.301</td>
<td>0.273</td>
<td>0.189</td>
<td>0.123</td>
<td>0.395</td>
<td>1</td>
<td>0.738</td>
</tr>
</tbody>
</table>

By comparing correlation coefficients among the latent variables squared with the AVE indicators, we confirm the discriminant validity of the model.

5.2.3 Discussion and Analysis of Results

Table 4 displays the results of the research hypotheses after a bootstrap test is performed. We conclude that H1, H2, and H3 are valid hypotheses, while H4 and H5 are not.

Table 4: Hypotheses testing results

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Constructs</th>
<th>Path</th>
<th>Standard deviation</th>
<th>T-stats</th>
<th>P-value</th>
<th>Validated</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H1)</td>
<td>TUSE -&gt; TSKILL</td>
<td>0.395</td>
<td>0.11</td>
<td>3.573</td>
<td>0.000</td>
<td>Yes</td>
</tr>
<tr>
<td>(H2)</td>
<td>DUSE -&gt; SSKILL</td>
<td>0.357</td>
<td>0.112</td>
<td>3.172</td>
<td>0.002</td>
<td>Yes</td>
</tr>
<tr>
<td>(H3)</td>
<td>MFB -&gt; FUT</td>
<td>0.593</td>
<td>0.096</td>
<td>6.172</td>
<td>0.000</td>
<td>Yes</td>
</tr>
<tr>
<td>(H4)</td>
<td>TSKILL -&gt; FUT</td>
<td>0.094</td>
<td>0.165</td>
<td>0.571</td>
<td>0.568</td>
<td>No</td>
</tr>
<tr>
<td>(H5)</td>
<td>SSKILL -&gt; FUT</td>
<td>0.132</td>
<td>0.143</td>
<td>0.927</td>
<td>0.354</td>
<td>No</td>
</tr>
</tbody>
</table>

The results indicate that both design mentorship and technical mentorship have positive impacts on the development of soft skills and technical skills, respectively. Additionally, mentor responsiveness has a significant positive impact on the effectiveness of ChatGPT adoption as a mentor for future computing projects.

5.2.4 Qualitative Analysis

The open-ended question reveals insights into why H4 and H5 hypotheses are not validated. Student feedback on ChatGPT shows two key observations. First, many students find ChatGPT helpful for the technical and design aspects of their projects. Some even compare it favorably to existing resources, citing more accurate answers. ChatGPT aids with brainstorming and analysis options, positively impacting work quality. Second, some students express caution about using ChatGPT. They mention unreliability in coding and faulty code returns. Students emphasize the need to cross-reference ChatGPT's responses to ensure reliability. Despite skill improvement, code errors affect technical mentorship quality. The comments demonstrate that ChatGPT has helped students to progress in their work and improve their work quality. However, caution needs to be exercised while using ChatGPT as some students have pointed out its limitations, particularly in areas such as coding, where it can be highly inaccurate and return faulty code. Future research will explore other factors such as ease of use, trust, and ethics to better understand ChatGPT's impact on various computing activities. Expanding the sample size to include multiple computing courses would offer a more comprehensive understanding of ChatGPT adoption in diverse contexts. Techniques like focus group interviews and longitudinal studies could further enhance data collection.

6. Conclusion

Our research studies ChatGPT's impact as a mentor in undergraduate computing projects. Mentor responsiveness emerges as a crucial factor influencing the future adoption of AI tools. Providing responsive mentorship with technical and design support is key to AI tool adoption.
Our study offers valuable insights into factors influencing ChatGPT’s adoption as a mentor and provides practical implications for educators integrating AI into computing projects.

References


Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics.


Deep Knowledge Tracing is an implicit dynamic multidimensional item response theory model

Jill-Jênn VIE\textsuperscript{a}, Hisashi KASHIMA\textsuperscript{b}
\textsuperscript{a}Soda Team, Inria, France  
\textsuperscript{b}Kyoto University, Japan  
* jill-jenn.vie@inria.fr

Abstract: Knowledge tracing consists in predicting the performance of some students on new questions given their performance on previous questions, and can be a prior step to optimizing assessment and learning. Deep knowledge tracing (DKT) is a competitive model for knowledge tracing relying on recurrent neural networks, even if some simpler models may match its performance. However, little is known about why DKT works so well. In this paper, we frame deep knowledge tracing as a encoder-decoder architecture. This viewpoint not only allows us to propose better models in terms of performance, simplicity or expressivity but also opens up promising avenues for future research directions. In particular, we show on several small and large datasets that a simpler decoder, with possibly fewer parameters than the one used by DKT, can predict student performance better.

Keywords: Knowledge Tracing, Item Response Theory, Side Information.

1. Introduction

Adaptive testing and personalized learning are precious technologies that have been enabled by tracing the knowledge of previous students. They rely on a response model of the learners: if we know how people learned in the past, we can optimize testing or learning for new students. This is the model-based approach of reinforcement learning, commonly referred to as knowledge tracing in the educational data mining literature.

Formally, knowledge tracing relies in predicting the outcomes of students on some items, given their previous interactions on some items. Numerous models have been proposed for knowledge tracing such as deep knowledge tracing (DKT) proposed by Piech et al. (2015) that relies on a recurrent neural network (RNN). However, Wilson et al. (2016) have matched the performance of DKT with a unidimensional item response theory (IRT) model that can be seen as online logistic regression. In order to advance the field, we need to understand what distinguishes these models, and how to build upon them. For example, we can notice that these models do not rely on the exact same data. IRT usually learns a difficulty parameter per item while DKT were initially strictly used on skill data for the sake of the comparison with Bayesian knowledge tracing (BKT).

In this paper, we show how we can bridge both categories of models using an encoder-decoder architecture. These architectures are usually encountered in sequence-to-sequence scenarios such as machine translation (Cho et al., 2014) and got popular with the rise of transformers (Vaswani et al., 2017). We open the “black box” of DKT and explain how it relates to existing, well known models.

Our main contribution is an encoder-decoder architecture that takes DKT, IRT and other models such as Performance Factor Analysis (PFA) as special cases. We demonstrate using several real datasets that it seems better to learn unidimensional parameters for item, skills, which is not what the vanilla DKT model does, relying instead on multidimensional parameters for skills. In this paper, we are not so much interested in state-of-the-art performance but in understanding how models are related to each other and how to interpret
their parameters, so that they can also be visualized. We notably show that metadata considered can have a considerable impact on performance.

We first expose related work, then our approach: summarize & predict. We subsequently detail our datasets, experiments, and see the influence of specific components of our architecture in the results.

2. Background and related work

Formally, knowledge tracing can be defined as follows. Let us denote $I$ the set of items of some test, and $K$ the number of knowledge components assessed. For each student, at each time step $t$, we know the sequence of items and outcomes already given to the student $(q_1, a_1, \ldots, q_t, a_t)$ (here, time $t$ is excluded as it is what we want to predict) where $q_\tau \in I$ and $a_\tau \in \{0, 1\}$. We may as well have some side information about $q_t$ such as the knowledge components (KCs) required by $q$, denoted $KC(q_t) \subseteq \mathcal{P}\{1, \ldots, K\}$, sometimes represented by a binary vector from $\{0, 1\}^K$ in a so-called $Q$-matrix that indicates which knowledge components are assessed. Using this information, we need to predict outcome $a_t$ using a model that estimates a probability $p_t$. We will note $\sigma$ the sigmoid function: $\sigma : x \mapsto \frac{1}{1+\exp(-x)}$. We now describe some models for knowledge tracing, of two different kinds.

2.1. Sequence-based architectures

**Deep knowledge tracing models.** Deep knowledge tracing (DKT) is usually described as a black-box model that takes into input pairs $(q_1, a_1, \ldots, q_t, a_t)$ from some student, and outputs a vector of probabilities $y$ such that $y_k$ is the probability that the student will answer correctly an item requiring skill $k$. More precisely, in a DKT model, the actual probability that a certain user correctly answers an item that requires skill $k$ at time $t$ (event $R_{tk} = 1$) is given by:

$$Pr(R_{tk} = 1) = \sigma(\langle h_t, v_k \rangle + w_k)$$

where $v_k \in \mathbb{R}^d$, $w_k \in \mathbb{R}$ is a representation of the skill $k$ learned by DKT that does not evolve over time and $h_t = \text{LSTM}[h_{t-1}, q_t, a_t]$ if $t > 1$, $h_1 = 0$ otherwise, is a representation of the user that evolves over time. Montero et al. (2018) have shown that DKT can trace knowledge efficiently even when skills are interleaved, and that it shares information between skills. In the original DKT paper, Piech et al. (2015) assume that each item is only related to one skill among $1, \ldots, K$. They learn a joint representation in $\{0, 1\}^2K$ for the pair $(s, a)$ where $s = KC[q_i]$, because they claim that separate representations for $s_i$ and $a_i$ degraded performance. When there are too many pairs $(s, a_i)$, they use fixed low-dimensional representations instead of $\{0, 1\}^2K$, which can be seen as setting a random embedding for each pair. It makes sense to use skill instead of items for encoding the actions of the student, in order to have redundancy and avoid the item cold-start problem. The drawback is that in their vanilla form, DKTs cannot handle multiple skills per item. In this paper, we show in particular how multiple skills can be taken into account, which can be used to improve DKT.

In DKT-DSC, an extension of DKT, Minn et al. (2018) use an encoding of the triplet $(s, a, c)$ as input metadata where $c_i$ is a dynamic clustering information of the student based on a vector counting successful and unsuccessful attempts, updated at each time step. They managed to outperform DKT on several datasets.

Ding & Larson (2019) have shown that the encoder of DKT with untrained, random weights, is a strong baseline; which seems to indicate that the latent state $h_t$ can be a random embedding of the sequence while the final fully connected layer is mainly responsible for the good predictive performance.
More recently, transformers have been proposed for knowledge tracing (Pandey & Karypis, 2019; Ghosh, Lan, & Heffernan, 2020). Self-attention can be used to model how the samples we want to predict relate to previous observations from the student.

2.2. Non-sequence-based architectures

Multidimensional item response theory (MIRT). In item response theory, one usually does not assume that the examinee’s ability evolves over time. The probability that user $i$ correctly answers item $j$ (event $S_{ij} = 1$) is:

$$
\Pr(S_{ij} = 1) = \sigma(\langle \theta_i, v_j \rangle + w_j)
$$

where $\theta_i \in \mathbb{R}^d$ is a learned representation of user $i$ called multidimensional ability, $v_j \in \mathbb{R}^d$ is a learned representation of item $j$ called discrimination parameter and $w_j \in \mathbb{R}$ is a bias parameter representing the easiness of item $j$. This model is popular in the psychometric literature, because it can enable computerized adaptive testing: adapting the assessment, given the currently estimated performance of the student. It is indeed possible to make the model dynamic by updating the maximum likelihood estimate (MLE) of $\theta_i$ at each new observed point, which can take time to estimate. This is exactly what Wilson et al. (2016) do, to demonstrate a better performance of updated IRT compared to DKT. Maximum likelihood estimation (MLE) is possible with simple models such as IRT, but for more complex architectures or priors, it is not always feasible because it increases time complexity during the test phase. This is why simpler online updates have been proposed such as multivariate Elo (Abdi et al., 2019) to train dynamic MIRT models.

Performance Factor Analysis and count-based models. Several models rely on attempt counts, such as Additive Factor Model (AFM) and Performance Factor Analysis (PFA) proposed by Pavlik, Cen, & Koedinger (2009). The estimated ability of the student at time $t$ is a linear function of their number of previous successful $W^{1:t}_k$ and unsuccessful attempts $F^{1:t}_k$ over skill $k$ up to time $t$ (again, time $t$ information is excluded, as it is not observed yet at prediction time). An important remark is that here, we forget the order in which the items were solved, we are considering mere counts. DAS3H proposed by Choffin et al. (2019) maintains counters in several time windows, which enables the modeling of forgetting. In PFA, $p_t$ depends on a weighted sum of the counts, according to the knowledge components that are assessed at time $t$ and specified in $KC(q_t)$:

$$
D_{PFA} : p_t = \sum_{k \in KC(q_t)} \beta_k + \gamma_k W^{1:t}_k + \delta_k F^{1:t}_k
$$

where parameters $w = [\beta, \gamma, \delta]$ are learned, respectively scalar parameters for skills, wins, fails. Please note that this formulation can handle multiple skills.

Knowledge Tracing Machines. Vie & Kashima (2019) have shown that it is possible to learn representations for all users, items, skills in a test and combine them in a pairwise manner. Interestingly, most existing models for knowledge tracing such as IRT, PFA, MIRT are special cases of KTMs, according to the representations considered in the modeling. In this paper we want to emphasize the importance of metadata. Gervet et al. (2020) have also shown that predicting performance at the item level instead of at the skill level could improve performance of DKT.

3. Our approach: summarize & predict

We see all these models through the same lens: our architecture relies on two main components, an encoder $E$ that summarizes prior data and a decoder $D$ that predicts student performance; both are trained jointly. They respectively require an expression of the metadata of former actions from the student $(q_{1:t}, a_{1:t})$ and metadata of assessment at time $t$: which item is being attempted, which skills are being assessed, etc.

$$
\begin{align*}
&h_t = E((q_{1:t}, a_{1:t})) \\
p_t = D(h_t, q_t)
\end{align*}
$$

$t = 1, \ldots, T$
In particular, $h_1 = E(\emptyset) = 0$ where $\emptyset$ denotes the empty sequence. In this expression, $h_t$ represents the learned representation of the user at time $t$, and $p$, the probability that the attempt of the student at time $t$ will be correct. Please note that this modeling is rich enough to encompass broader tasks than knowledge tracing, for example we can model the event that some user watched the video of some lesson, with the corresponding metadata. If we can take advantage of this kind of data, we can hopefully recommend new videos to some student to optimize their learning.

3.1 Summarize: encoding a dynamic representation of what happened

The encoder $E$ is a function of past history. It can either be simple counts to remember how many times a student succeeded or failed an attempt at a single skill, or a more complex function like a RNN in the case of DKT: $E_{DKT}: h_t = RNN(h_{t-1}, q_t, a_t)$ if $t > 1$, $h_1 = 0$ otherwise. In order to share information across items, it is common to use pairs $(s_t, a_t)$ as sequence inputs where $KC'(q_t) = \{s_t\}$. If $q_t$ is associated with multiple KCs, some tricks consist in taking one at random, or creating a new skill that represents the combination of multiple skills.

3.2 Predict: combine the learned representations to make a prediction

The objective of the decoder $D$ is to combine the learned representation $h_t$ of the student at time $t$ with some parameters involved in the assessment at time $t$ such as which item was administered, which skills were assessed, etc.

In order to build a decoder we need to decide two main characteristics: what is the metadata taken into account for the prediction (item $q_t$, skills $KC'(q_t)$, wins $W_k^{1:t}$, fails $F_k^{1:t}$), and what is the embedding size of those features.

3.2.1 Output metadata

If the output is solely based on item ID, we get a model like IRT or MIRT according to the embedding size of items. If the output is based on skill IDs $KC(q_t)$, number of successful attempts (wins $W_k^{1:t}$), number of unsuccessful attempts (fails $F_k^{1:t}$), and the embedding size is 1, we get a model similar to PFA.

3.2.2 Decoder of same embedding size as $h_t$

As an example, MIRT computes a dot product of the user embedding (usually denoted as $\theta_t$ in the literature, here $h_t$ for consistency) and the embedding of item $q_t$ (usually called discrimination, here $v_{q_t}$), plus a bias representing the easiness of the item (here, $w_{q_t}$):

$$D_{MIRT}: p_t = \sigma(h_t, v_{q_t} + w_{q_t})$$

The vanilla DKT model has a fully connected layer that converts the latent state $h_t$ to the prediction of performance of every possible skill. In practice, we are only interested in the skill $s_t = KC(q_t)$ assessed at time $t$ (assumed single by DKT) and DKT implicitly computes a dot product of the user embedding $h_t$ and the skill embedding $v_{s_t}$, leading to the following expression:

$$D_{DKT}: p_t = \sigma(h_t, v_{s_t} + w_{s_t})$$

Now, what should strike the reader is that the expressions of $D_{MIRT}$ and $D_{DKT}$ are almost the same: only the metadata is different, item ID for MIRT and skill ID for DKT. This is why DKT, in its item output form, can be seen as a dynamic MIRT model, where a RNN looks at the sequence of observations to estimate ability parameters at current time. See Figure 1 for an illustration.

In fact, Gervet et al. (2020) have shown an improvement in performance when considering items as output of DKT instead of skills as output.
3.2.3 Decoder of different embedding size than encoder

The encoder usually needs a high dimensional embedding to encode learning dynamics, notably if it is a RNN. The item and skill embeddings do not need to be the same dimension as the latent state $h_t$. Indeed, in practice, most hard baselines for knowledge tracing such as PFA only model skill parameters with one dimension. This is why we suggest reducing the dimension of $h_t \in \mathbb{R}^d$ before making the predictions, using a fully connected layer $d \rightarrow d'$:

$$h'_t = Ah_t + b$$

where parameters $A \in \mathbb{R}^{d' \times d}$ and $b \in \mathbb{R}^{d'}$ are learned. In the most extreme example, $h'_t$ can be scalar, i.e. unidimensional, $d' = 1$. The advantage is that it drastically reduces the number of parameters. For example, a decoder of embedding size 1 using (item, skill, wins, fails) as metadata would be a blend of IRT and PFA models and would look like this:

$$D_{swf}: p_t = \sigma \left( h'_t + w_{q_t} + \sum_{k \in KC(q_t)} \beta_k \text{skill} + \gamma_k W_{k}^{1:t} + \delta_k F_{k}^{1:t} \right)$$

where weights $w$ are learned for each item and weights $\beta, \gamma, \delta$ are learned for each skill. We will denote such a decoder as “iswf $d' = 1$” later on. Another advantage of this decoder compared to vanilla DKT is that it can consider multiple skills associated to $q_t$. To see how embeddings of higher dimension $d' > 1$ can be considered for multiple metadata, see Vie & Kashima (2019).

3.3 Evaluation

We optimize the log-loss, also known as mean negative log-likelihood:

$$L(a, p) = \sum_{i=1}^{n} \log \left( 1 - |a_i - p_i| \right).$$

This log-loss can be computed for a batch of students, but we have to take into account that all students did not attempt the same number of items, so student sequences within a batch may have different lengths.

Once the parameters of the encoder $E$ and the decoder $D$ have been trained, for a new student it is easy to unroll the encoder on the sequence of outcomes $(q_t, a_t)$ encoded as input metadata, get the corresponding latent states $h_t$, and feed them to the decoder with the output metadata to get the predictions, and compute performance metrics.

We report all results according to their accuracy (ACC) and area under the curve (AUC).
4. Experiments

The code of our experiments is on GitHub: https://github.com/jilljenn/dktm.

4.1 Models

We tried different combinations of encoders and decoders. As sequence-based encoder, we used a RNN of latent dimensionality $d$ where the encoding of actions $\{q_i, a_i\}$ is sampled from a Gaussian multivariate distribution of dimension $d$ and fixed. We used GRU (Gated Recurrent Unit) because it has fewer parameters than LSTM (Long Short-Term Memory), so it is less prone to overfitting (Chung et al., 2015). In the reported results, when encoder is listed as “none”, it means only the wins and fails parameters $W_k^{1,1}$ and $F_k^{1,1}$ were counted. Decoders are described by an $n$-gram among letters “swf”, respectively item, skill, wins and fails; and an embedding size $d'$. Examples of decoders are given in the previous section. The corresponding baselines are:

- **DKT**, which has a RNN as encoder and multidimensional embeddings within its decoder. The decoder can either be “s” if DKT makes predictions at skill level or “i” if DKT makes predictions at item level.
- **PFA** just keeps track of successful and unsuccessful attempts so there is no sequence-based encoder; important metadata is skill, wins, fails so “swf” and the embedding size is 1 as scalar parameters are learned.
- **Decoders “swf $d' = 1$” (PFA) and “iswf $d' = 1$” are particular cases of logistic regression (LR) with sparse features.**

4.2 Datasets

We tried our approach on the following datasets that have diverse characteristics.

**Fraction.** 536 middle-school students attempting 20 fraction subtraction exercises requiring 8 KCs such as, being able to put fractions at the same denominator. All students attempted all items, so this dataset is fully specified. The dataset and description of KCs can be seen in (DeCarlo, 2010). This dataset is particularly interesting because it is small scale, so complex models may overfit. Also, most items measure several KCs at the same time. As we have access to the full response pattern of all students, we assume that items are solved from item 1 to item 20.

**Assisments 2009.** 346,860 attempts of 4,217 students over 26,688 math items requiring 123 KCs (Feng et al., 2019). Some of the items require multiple KCs, up to 4. Students attempted between 1 and 1,382 exercises.

**Berkeley.** 562,201 attempts of 1,730 students over 234 CS-related items in a MOOC provided by Berkeley. Items can be of 29 categories, that we used as KCs. Each item belongs to a single category. Students attempted between 2 and 1,020 exercises. This dataset is private.

4.3 Data preparation

We used 5-fold cross-validation: we split each dataset into 5 folds, predict any of them using the remaining ones, and average the results. In practice there are many challenges, fortunately already known in the natural language processing community.

**Batching.** Whenever there are too many data points, in order to save memory we need to sample a batch of users. So the log-loss is computed on a batch.

**Sequences of uneven lengths.** Within a batch, different users may have attempted a different number of questions. However for efficiency, it is better to vectorize as much as possible and compute the log-loss over matrices. So we use a mask to know where the sequences end for each student.

**Long sequences.** On long sequences, DKT is hard to train. First, it may forget information on a long sequence, second the computation of the gradient may take time,
vanish to zero or conversely get arbitrary high. As a remedy, we train on windows of fixed size called BPTT: backpropagating through time (Werbos et al., 1990). The latent state of the previous batch is fed to the next batch.

4.4 Implementation details

We optimize our model using the Adam optimizer (Kingma & Ba, 2014) with learning rate γ = 0.005 and weight decay λ = 0.0005 (equivalent to $L_2$ regularization). We train using 100 minibatches and use time windows of size 100. For the RNN, we used GRU with one layer, no dropout and latent dimensionality 2 or 50. Our implementation is in PyTorch. We ran the Fraction dataset experiments on CPU and the Assimtments and Berkeley dataset experiments on GPU. Training was stopped after 200 epochs.

5. Results and Discussion

Results are shown in Tables 1 to 3 where the best models are shown in bold.

Table 1. Results on the Fraction Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Encoder</th>
<th>Decoder</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>GRU $d = 2$</td>
<td>iswf $d' = 1$</td>
<td><strong>0.880</strong></td>
<td><strong>0.944</strong></td>
</tr>
<tr>
<td>LR</td>
<td>None</td>
<td>iswf $d' = 1$</td>
<td>0.853</td>
<td>0.918</td>
</tr>
<tr>
<td>PFA</td>
<td>None</td>
<td>swf $d' = 1$</td>
<td>0.854</td>
<td>0.917</td>
</tr>
<tr>
<td>DKT</td>
<td>GRU $d = 2$</td>
<td>$i d' = 2$</td>
<td>0.772</td>
<td>0.844</td>
</tr>
<tr>
<td>DKT</td>
<td>GRU $d = 2$</td>
<td>$s d' = 1$</td>
<td>0.761</td>
<td>0.839</td>
</tr>
</tbody>
</table>

Table 2. Results on the Assimtments Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Encoder</th>
<th>Decoder</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>None</td>
<td>iswf $d' = 1$</td>
<td><strong>0.714</strong></td>
<td><strong>0.748</strong></td>
</tr>
<tr>
<td>Ours</td>
<td>GRU $d = 50$</td>
<td>iswf $d' = 1$</td>
<td>0.711</td>
<td>0.726</td>
</tr>
<tr>
<td>DKT</td>
<td>GRU $d = 50$</td>
<td>$i d' = 50$</td>
<td>0.691</td>
<td>0.701</td>
</tr>
<tr>
<td>PFA</td>
<td>None</td>
<td>swf $d' = 1$</td>
<td>0.682</td>
<td>0.686</td>
</tr>
</tbody>
</table>

Table 3. Results on the Berkeley Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Encoder</th>
<th>Decoder</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>GRU $d = 50$</td>
<td>iswf $d' = 1$</td>
<td><strong>0.707</strong></td>
<td><strong>0.778</strong></td>
</tr>
<tr>
<td>LR</td>
<td>None</td>
<td>iswf $d' = 1$</td>
<td><strong>0.704</strong></td>
<td><strong>0.775</strong></td>
</tr>
<tr>
<td>DKT</td>
<td>GRU $d = 50$</td>
<td>$i d' = 50$</td>
<td>0.684</td>
<td>0.751</td>
</tr>
<tr>
<td>PFA</td>
<td>None</td>
<td>swf $d' = 1$</td>
<td>0.630</td>
<td>0.683</td>
</tr>
</tbody>
</table>

For all datasets, we discovered models that outperformed the vanilla DKT model. More precisely, decoders of embedding size $d' = 1$ outperform decoders of higher embedding size. This seems to indicate that, although sequence-based encoders (GRU) may be better in some cases to learn the dynamics of student knowledge, it is always better to represent assessment information (items, skills) as scalars, to avoid overfitting. Besides improving the performance, reducing parameters induces a speedup in training and allows interpretability.

On the large datasets Assimtments and Berkeley, logistic regression, that uses a simple counter of successful and unsuccessful attempts at skill level, is among the top models. It seems to indicate that for big datasets, logistic regression may be enough, because the number of prior successes and failures is enough to predict. What is even more surprising, is that on the small dataset Fraction, the top performing model uses a recurrent neural network as encoder. It may be because the sequences are small: every student only attempts 20 items. So the sequences of successes and failures make students more unique than their counters of successes, inducing a more personalized estimation.

Considering an item bias does not improve a lot the quality of the predictions on the Fraction dataset, maybe because the fraction subtraction task is particularly easy to describe.
using KCs, so the KCs are enough to characterize the items. However, on AssiStments, one can see the improvement when considering an item bias, see the difference between PFA and LR (0.06 improvement in AUC). Same goes for Berkeley (0.09 improvement in AUC), where the KCs are actually mere categories of items.

6. Conclusion

Most models in the educational data mining literature model users, items and skills with multidimensional parameters. In this study though, we showed that while it is indeed important to model the dynamics of the students with multidimensional parameters, the items do not necessarily need to be multidimensional to come up with really strong models. In particular, logistic regression is among the top models. This is encouraging because we come up with better models that have fewer parameters. DKT may be slow to train not because of the recurrent neural network but because of the fully connected layer at the output, that contains the most parameters. In this paper, we suggest to replace it with a lower-dimensional layer.

As future work, we plan to try to use other kinds of side information. For example, for the Duolingo SLAM dataset (Settles et al., 2018) for second language acquisition modeling, contestants had to predict whether a learner would get a word correct, and the best approaches were combinations of DKT with word embeddings (Mikolov et al., 2013). Using word content, be it only n-gram features, allows to learn what phonemes are harder for which categories of people. Also we will try to encode richer actions for the encoder, because it is costly for humans to define the knowledge components related to all items, and deep learning models may be able to recover most of those with little supervision, reducing the workload on practitioners.

7. Acknowledgements

We thank Samuel Girard for his insightful comments. This work has been supported by JST CREST Grant Number JPMJCR21D1.

References


Immediate Feedback in Computational Thinking: Generating hints using a Knowledge Graph

Nitesh Kumar JHA\textsuperscript{a}, Plaban Kumar BHOWMIK\textsuperscript{b} & Kaushal Kumar BHAGAT\textsuperscript{a}*

\textsuperscript{a}Advanced Technology and Development Centre, Indian Institute of Technology Kharagpur, India
\textsuperscript{b}G.S Sanyal School of Telecommunications Indian Institute of Technology Kharagpur, India
*kkntnu@hotmail.com

Abstract: In this study, we present an Online Inquiry-based Learning Platform for Computational Thinking (CT-ONLINQ) to develop the CT skills of students. The platform provides immediate feedback with hints to support students during problem-solving activities and encourages them to explore multiple solutions for a problem. The hints are generated using a Knowledge Graph that stores information about the solution details of a problem. A six-week study was conducted on 79 high school students to determine the effectiveness of the platform with hints on students' CT skills. The results showed that the students improved their CT skills significantly after the intervention. Findings suggest that the CT-ONLINQ platform consisting of IBL-based CT activities with immediate feedback could help school students improve their CT skills.

Keywords: Computational Thinking, Immediate feedback, Knowledge Graph, Inquiry-based learning

1. Introduction

Computational Thinking (CT) is a cognitive ability that allows individuals to develop computational solutions for a variety of problems (Wing, 2006). Inquiry-Based Learning (IBL) is an approach to learning that guides students through various phases, starting with exploring a problem, followed by collecting data to develop models, generating findings, and rigorously evaluating those findings to gain new understanding (Pedaste et al., 2015). In terms of the learning process, IBL in science and CT education are relatively similar due to the cyclic process of repeated revisions and refinement (Hoppe and Werneburg, 2019).

The current CT research focuses on teaching coding concepts to students (Jacob et al., 2020), but researchers argue that problem-solving, creativity, and algorithmic thinking should be taught, as these transferable skills hold greater value (García Peñalvo et al., 2016). In addition, studies on IBL in CT concentrate on creating logical artifacts and testing them to improve CT abilities (Hoppe and Werneburg, 2019). Moreover, researchers argue that analysis and evaluation of artefacts is critical to improving learning outcomes (Prayogi and Yuanita, 2018). To overcome the above issues, we have created CT-ONLINQ, an online platform for Computational Thinking (CT) that follows IBL-based CT activities (Jha et al., 2023). The platform provides hints and prompts during problem-solving activities, which are stored in a Knowledge Graph (KG). In KG, concepts are represented as nodes, and relationships between concepts are represented as edges. Different forms of immediate feedback have been developed by researchers to support coding and CT activities (Basu et al., 2017). However, there are lack of studies that provides hints stored in KG during IBL-based CT activities. The current study discusses the design of KG and evaluates the effect of the platform CT-ONLINQ on CT skills of high school students. The current study addresses the following research question (RQ):

- RQ: Did the online platform CT-ONLINQ with immediate feedback improved CT skills of high school students?
2. The CT-ONLINQ platform

CT-ONLINQ is an online educational platform that supports IBL-based CT activities (Jha et al., 2023). A brief description of the platform architecture and example interface tab is shown in Jha et al. (2023). The platform uses a KG to store hints and answers during IBL-based CT activities. Figure 1 shows a basic schema structure of the KG with properties for each problem. The KG schema is divided into three subschemas: Answer comparison, Goal, Hints and Answers, as described below.

- **Answer comparison**: It contains problem’s question and the sequence of goals related to CT step.
- **Goal**: It contains step name, task statement, the best solution, hints and answers sequence related to the task, and a no matched answer section that contains hints when input does not match with any correct or incorrect answer.
- **Hints and Answers**: It contains type of answer (correct/incorrect) that matches the input, an alternate answer to compare with the input, a prompt for the hint, and three types of problem-specific hints: basic (with little elaboration), intermediate (with more elaboration), and alt hint (for alternate correct answer). The alternate correct answer (only available for the correct answer type) is a solution but not the best solution for the task.

![Figure 1. Schema of the Knowledge Graph](image)

The platform parses the graph to provide relevant hints based on the number of attempts made during problem-solving activities. For example, Consider the problem statement ‘Write an algorithm to find the sum of numbers 1+2+…+60’. For an incorrect answer in decomposition step- ‘Find a way to add number from 1 to 60’ the hints are: basic(attempts<4)- can you do better? Add number pairwise to find total sum; intermediate (attempts<6)- can you do better? Find pairs having sum 61; Advanced- (1+60)+(2+59)+…(30+31).

3. Methodology

The experiment involved 79 high school students, aged 13-15, from a public high school in India. All the students were given immediate feedback with hints during problem-solving activities. We adapted two different sets of 12 questions (max score 108, pre and post-test) from Bebras Thinking Challenge (Bebras-Ireland, 2020). Initially, the students signed up to the CT-ONLINQ platform, completed an example activity (see Jha et al., 2023), and completed bebras questions (pre-test with Cronbach’s alpha 0.74). Next, the students completed six
problem-solving activities on the platform and completed post-test with Cronbach’s alpha 0.77.

4. Result and Discussion

The average pre-test score was 42.11 with a standard deviation (SD) of 10.69, while the post-test score was 55.43 with a SD of 14.07. We conducted a paired samples t-test to compare the pre and post-test scores to assess the improvement in CT skills. Findings showed that there was a significant difference in CT skills ($t=6.697$, $p=.000$, $d=1.06$), with an average difference of 13.316 (SD=1.988). The result supports the study that useful feedback and hints improved student understanding in a CT-based learning environment (Basu et al., 2017).

5. Conclusion and future work

The present study investigated the effect of CT-ONLINQ with immediate feedback on the CT skills of high school students. A total of 79 students from high school participated in the current study. The platform implemented the IBL-based CT workflow, allowing students to develop algorithmic solutions, analyze the algorithms, and compare the algorithms. The platform provided immediate feedback at each CT step. Hints and answers were stored in the KG. The findings confirmed that the intervention significantly improved the CT skills of students working on the platform.

With regards to future work, the current study examined the effect of IBL-based CT activities on the online platform but did not compare them with offline IBL-based CT activities. Future studies can compare the effects of offline versus online IBL-based CT activities on the enhancement of CT skills. The study provides a new way to teach CT skills to high school students using IBL-based CT activities without any coding practice.

References


Composite Score for ChatGPT Prompt Efficiency: A Computational Linguistic Analysis of Engineered Chatbot Prompts

Michelle BANAWAN
Asian Institute of Management, Philippines
mbanawan@aim.edu

Abstract: The use of chatbots has become increasingly popular in the field of education. Hence, the quality of the prompts used by chatbots can greatly influence the quality of the generated responses which in turn contributes to desirable outcomes related to the overall learning experience. However, even as known criteria for efficient prompts are prevalent, there is a dearth of measurable and concrete linguistic factors that guide prompt engineering. This paper presents a composite score for ChatGPT prompts which can be made applicable to other foundational generative AI chatbots. Through a computational linguistic analysis of known efficient prompts used in learning, emergent linguistic factors point to the relationship of linguistic features and the confidence of ChatGPT responses to well-structured prompts that use the said linguistic features. The linguistic features are the average collostructural strength, collostructural ratio diversity, specificity, and academic language use. These features depict the quality of prompts that pertain to the grammatical structure, specificity, and relevance to the task at hand, and academic language use. Further, these features constitute a composite score for prompts introduced in this study that represent linguistic efficiency and subsequently, correlates to perplexity or certainty estimates of the generated responses of ChatGPT.

Keywords: ChatGPT, PCA, NLP, computational linguistics, prompt engineering

1. Introduction

   Conversational agents such as chatbots are rapidly emerging as a promising technology for education and learning. With the help of advanced language models, chatbots can facilitate personalized learning (Allen, et al., 2016; Ashok, et al., 2021; Holotescu, 2016), provide instant feedback (Graesser, 2015; McCarthy, et al., 2022, and engage students in an interactive and engaging manner (Lin & Chang, 2020). However, the effectiveness of these chatbots, specifically generative AI-based chatbots, heavily relies on the quality of prompts that they use to elicit responses from the AI large language models (LLMs) (Dang, et al., 2022).

   The challenge lies in engineering prompts that can generate accurate and comprehensive responses from LLMs. While many existing prompts are effective, they often rely on domain-specific knowledge, which limits their generalizability across different educational settings. Moreover, there is a lack of understanding of the linguistic features that contribute to the effectiveness of prompts, which makes it difficult to develop effective prompts systematically.

   The current state of prompt engineering for educational chatbots is largely characterized by a trial-and-error approach, where developers often rely on ad-hoc strategies to generate prompts. This approach is not only time-consuming and resource-intensive but also limits the effectiveness of the chatbots in facilitating learning.
To address these challenges, this paper presents a computational linguistic analysis of prompts which can be leveraged in enhancing the learning experience. By leveraging computational methods, this work attempts to quantify prompt efficiency.

2. ChatGPT\(^1\) and Generative AI Tools

Generative language models have revolutionized the field of natural language processing (NLP), enabling the creation of chatbots that can generate human-like responses to user inputs. Among these models, GPT (Generative Pre-trained Transformer) stands out as one of the most widely used and effective models for generating text.

The GPT model is based on the transformer architecture, a deep neural network that is capable of processing input sequences of variable length and generating output sequences of variable length. The model is trained on a large corpus of text data, and its parameters are optimized to minimize the prediction error of the next token in the sequence given the previous tokens. One of the most well-known implementations of the GPT model is ChatGPT, a conversational agent that can generate human-like responses to user inputs. ChatGPT is based on the GPT-3, GPT-3.5, and now GPT-4 architecture, which is currently one of the largest and most powerful models available for generating text.

In the field of education, ChatGPT and other generative AI tools have numerous applications. One of the most promising applications is in the area of personalized learning. Chatbots can be used to deliver customized learning experiences, adapting to the needs and preferences of individual learners. They can also provide instant feedback and support, helping learners to overcome obstacles and stay motivated.

Another application of generative AI tools in education is in the creation of educational content. Chatbots can generate text-based content such as summaries, study guides, and practice questions, which can be used to supplement traditional educational materials. Additionally, they can be used to create interactive learning experiences, such as simulations and games, that engage learners in a fun and immersive way.

ChatGPT, as well as other generative AI tools, has now become prevalent in education and across many domains and industries. These tools have the potential to revolutionize the field of education by enabling personalized learning, creating interactive content, and providing instant feedback and support.

2.1 Prompt Engineering for Education

Effective prompt engineering is critical for generating accurate and comprehensive responses from LLMs in various domains, most especially in educational settings. Prompt engineering involves crafting well-designed prompts that clearly and concisely communicate the intended learning objectives and desired outcomes. However, prompt engineering is not a straightforward process, and often requires iterations of trial and error, feedback, and back-and-forth conversations with ChatGPT as additional information is required to achieve the desired information in generated responses. By investing in prompt engineering and creating effective prompts, educational institutions can harness the power of ChatGPT for many applications including to enhance student learning and automate time-consuming tasks, such as grading and feedback, to name a few. However, it's important to note that prompt engineering requires careful planning and ongoing refinement to ensure that the prompts are effective and aligned with the intended learning outcomes. As such, concrete measures of what constitute efficient prompts should be investigated and put in place.

Ongoing work related to prompt engineering of large language models (LLMs) have found that a number of strategies affect the quality of the generated outcome, e.g. repetition, providing examples, using code-like structures, or choosing the right instructions

---

\(^1\) https://openai.com/product/gpt-4
(Zamfirescu-Pereira, et al., 2023), focusing on the word order of prompts (Yao Lu, 2021), using structured prompt patterns that explicitly contain elements like intent, context, motivation (White, et al., 2023), to name a few. However, there is little work on probing into the linguistic features that characterize well-structured and efficient prompts that generate confident responses from LLMs and other generative AI models, like ChatGPT.

2.2 Research Question
The central aim of the study is to seek to develop a comprehensive understanding of the linguistic factors that contribute to the effectiveness of prompts for LLMs which can benefit educational settings. The research question guiding this study is: What are the linguistic features of efficient ChatGPT prompts? By answering this question, the study aims to articulate what specific linguistic markers should be present in formulating efficient prompts which would in turn generate more confident responses from LLMs and minimize the extent of hallucinations in the generated responses.

3. Methods
3.1 Prompt Collection
The corpus used in this study was collected from publicly available effective ChatGPT prompts. The sources include effective prompts for teachers from https://www.learnprompt.org/chat-gpt-prompts-for-teachers/ and used in curriculum management, classroom management, professional development, explaining difficult concepts and providing examples, as well as other use cases. Additional effective prompts were collected from https://github.com/f/awesome-chatgpt-prompts which is a github repository of effective ChatGPT prompts with 247 pull requests, 8.9k forks, and includes 2,190 downloads from its huggingface⁷ site as of March 2023. A total of 500 effective ChatGPT prompts (mean word count = 82.19; sd word count = 37.86) were gathered to comprise this study’s corpus.

3.2 Prompt Efficiency Evaluation
For this study, prompt efficiency is operationalized as the ability of a prompt to elicit coherent and confident responses usually characterized by having low perplexity scores. This metric is useful when the goal is to evaluate the effectiveness of prompts in generating high-quality responses from an AI generative model. The efficiency or quality of prompts was performed using the pipeline in figure 1. A random sample of 200 prompts from the corpus was used in calculating the perplexity scores of generated responses. Perplexity scores are used to determine the quality of generated responses given the prompts that are provided to a generative AI model. The lower the perplexity score, the more likely it is generated by ChatGPT (Malinka, et al., 2023) or an AI generative model. This metric has been used in works like determining if short texts are generated by a human or by ChatGPT(Yeadon, et al., 2023; Mitrović, 2023). Using open AI’s GPT3 API, the researcher generated responses to the sample corpus of 200 well-written prompts then approximated the perplexity scores of the generated answers using Hugging Face’s gpt models. The resulting perplexity scores of the generated results ranged from 9.84 to 52.1 with a mean of 28.17 (sd=13.4). The relatively low approximate perplexity score mean suggests that the prompts that were randomly sampled from the collected corpora used in the study were able to elicit generated responses that were relatively more certain. This could imply that the prompts are clear and effective, as the GPT model was able to generate coherent and well-formed text in response to them with relatively high certainty and, with post-hoc qualitative inspection, were observed to be more accurate.

⁷ https://huggingface.co/datasets/fka/awesome-chatgpt-prompts
3.3 Linguistic Index Extraction

Natural language processing is performed on the entire corpus of 500 prompts (see figure 2). Linguistic indices were extracted from the entire corpus of 500 prompts using the Suite of Analytic Linguistic Analysis Tools (SALAT; www.linguisticanalysistools.org). The SALAT NLP tools extract a wide array of linguistic indices that measure syntax, readability, lexical diversity, lexical sophistication, cohesion, grammar, spelling, and other structural and mechanics constructs. A total of 515 linguistic measures were extracted from the corpus of efficient prompts. To prepare the dataset for principal component analysis (PCA), the extracted indices were further preprocessed. First, indices with zero to only 5% variance were removed, as they do not provide significant information for the analysis. Next, highly correlated indices (≥ 0.799) were eliminated to avoid multicollinearity issues that may arise during PCA. Finally, any indices with missing or null values were removed. The resulting dataset contained 72 linguistic indices which were used for the PCA analysis. Prior to the PCA analysis, the dataset was standardized by scaling each index to have a mean of zero and a standard deviation of one.

3.4 Principal Component Analysis

After extracting the linguistic indices, we performed a principal component analysis (PCA) to identify underlying factors that explained the variability of the linguistic indices. We used the Scikit-learn library in Python to perform PCA on the dataset. The use of PCA in this study was chosen to provide a more comprehensive understanding of the linguistic patterns present in educational prompts, and for generating more targeted recommendations for prompt engineering in educational settings.

To ensure that the dataset was appropriate for PCA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was conducted. The KMO measures the proportion of variance among the observed variables that can be explained by the underlying factors. In this study, the overall KMO value was 0.65 and the individual features all had a KMO above 0.50, indicating that the dataset was suitable for PCA analysis. Moreover, a Bonferroni test of sample adequacy was also conducted to determine if the sample size was sufficient for PCA. The test indicated that the sample size was sufficient for PCA analysis (p < 0.05). A crucial step in PCA is to determine the number of principal components (PCs) to retain. The researcher used two methods for selecting the number of components: Parallel Analysis (PA) and Optimal Coordinates (OC). PA and OC are two methods for selecting the number of components in PCA that have advantages over the elbow method and eigenvalues > 1 method. Both methods consider the variability of eigenvalues and the correlations among
variables, resulting in more accurate and reliable estimates of the number of components to retain. The PCA analysis produced four principal components using computational methods (parallel analysis and optimal coordinates, see figure 4), explaining 60% of the total variance. The first component accounted for the highest proportion of the variance (22%), the second component accounted for 16% of the variance, while the last two components accounted for the lowest proportion (12% each). The components were then named and interpreted based on the linguistic attributes of the features that had the highest loadings (loadings threshold = 0.40).

Figure 4. Computational Solutions to Finding the Optimal Number of Components.

3.5 Qualitative Analysis of PCA
The researcher conducted a qualitative review of prompts that scored high and low in each of the emergent components. The primary objective of this step was to assist in interpreting and naming the emergent components and to scrutinize the specific prompts, thereby confirming the constitutive NLP indices that loaded onto each component. In this posthoc qualitative analysis the results of the PCA are studied in a more detailed and nuanced way to gain a better understanding of what each component represents in terms of the original variables.

3.6 Linguistic Components: Interpretation of PCA Results
The PCA analysis revealed a total of 4 components that accounted for up to 60% of the cumulated variance in the linguistic properties of efficient ChatGPT prompts (see table 1).

Table 1. Factor Loadings and Variances

<table>
<thead>
<tr>
<th>Component</th>
<th>Constitutive Indices (Original Variables with loading &gt; 0.40)</th>
<th>Proportion of Variance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic Association and Construction Strength</td>
<td>1. Average lemma construction combination frequency (academic, news, fiction, magazine)</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>2. Average approximate collostructional strength (academic, news, fiction, magazine)</td>
<td></td>
</tr>
<tr>
<td>Collostruction Ratio Diversity</td>
<td>1. Collostruction ratio ((academic, news, fiction, magazine)</td>
<td>16</td>
</tr>
<tr>
<td>Specificity</td>
<td>1. Sentence length</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2. T-unit length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Count of verb phrases</td>
<td></td>
</tr>
<tr>
<td>Academic Language Use</td>
<td>1. Frequency of academic words</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2. Combination of academic words</td>
<td></td>
</tr>
</tbody>
</table>

3.6.1.1. Component 1: Linguistic Association and Construction Strength (LACS)
The first component, which accounted for the largest amount of variance, 22%, is comprised of the average lemma construction combination frequency (ALCP) and the average approximate collostructional strength (AAC). The ALCP measures how frequently particular combinations of words are used together and how frequently particular combinations of words are used together. The AAC measures the strength of association between a given
word and the other words that typically co-occur with it. Both measures emerged specifically for words on specific genres, such as academic, news, fiction, and magazine and how they are combined in a prompt. Collostruction or a person’s collostructional preferences reveal interesting effects of the grammatical choice contributing to pragmatic phenomena, e.g., hedging or tentativeness (Rautionaho, 2021; Stefanowitsch & Gries, 2003). ALCP and AAC, refer to association strength which can be seen as an indicator of the level of lexical diversity, prototypicality, acquisition. Prototypicality allows for more natural and effective language. These features have been shown to be positively correlated with writing quality (Kyle, 2016). Thus, prompts that exhibit higher levels of association strength (ALCP and AAC) may be more effective in stimulating thoughtful and insightful responses from generative AI bots.

3.6.1.2. Component 2: Collostruction Ratio Diversity (CRD)
The collostruction ratio (CR) measures the degree of association between two words in each corpus. It is the ratio of the observed frequency of the collocate to the expected frequency of the collocate, given the frequency of the word. Component 2 focuses on the frequency of certain collocations within a prompt, namely the ratio of words that tend to co-occur with other words in specific domains, such as academic, news, magazine, fiction, and all domains. The CR is a measure of the strength of association between two words, and it has been found to be an effective indicator of semantic relatedness. An increase in collostruction ratio has been found to also depict an increase in writing quality (Mostafa & Crossley, 2020). CR Diversity has five constitutive indices, each of which is a CR for a different corpus: academic, news, magazine, fiction, and all. The CR for academic words measures the strength of association between the prompt and academic words, while the CR for news words measures the strength of association between the prompt and words commonly found in news articles, and so on. A prompt that has a diverse range of collocates (collostruction ratio) from different corpora may be more effective in eliciting a response from a chatbot, especially if the chatbot is designed to respond in a way that is relevant to the academic context which is intrinsically diverse in terms of topics and knowledge domains.

3.6.1.3. Component 3: Specificity
Component 3 is comprised of three constitutive indices: sentence length, unit length, and verb phrases. This component measures the degree of specificity and clarity of prompts. Sentence length is the average number of words per sentence in the prompt, which is a measure of syntactic complexity. Unit length is the average number of clauses per sentence, which indicates how much information is packed into each sentence. Verb phrases refer to the average number of verb phrases per sentence, which reflects the level of specificity and clarity in the prompt and are related to the writing quality (Mostafa & Crossley, 2020). Verbs play a critical role in conveying specific meaning and action in a sentence. Choosing the appropriate verb can enhance the clarity and specificity of the message being conveyed. Therefore, including verb phrases as a constituent index in the composite score can help assess the specificity and clarity of the language used in prompts, which can ultimately contribute to the effectiveness of the prompts in educational settings. Further, it is also notable that all three constitutive indices are indicators of syntactical complexity (Kyle & Crossley, 2018). While the syntax of a prompt does not necessarily determine the quality of the response generated by language models and chatbots. Both simpler and more sophisticated syntax can lead to high-quality responses, depending on the context and the specific task at hand. Ultimately, the syntax of a prompt should be chosen based on the intended audience and the specific goals of the task. Chatbots are capable of understanding and generating text with a wide range of syntactic structures.

3.6.1.4. Component 4: Academic Language Use (ALU)
Component 4, Academic Language Use, is characterized by a set of constituent indices that measure the frequency and combination of academic words in prompts. These indices include measures of average lemma frequency, construction frequency, and lemma-construction combination frequency of academic words. The use of academic language in
prompts is particularly important in educational settings, as it reflects the technical terms of academic subject areas and can help generative AI bots and language models extract domain-specific responses. Academic language is often characterized by its precision, specificity, and technical terminology. In academic writing, authors are expected to use language that is clear and unambiguous, and to avoid vague or imprecise expressions. As a result, academic language tends to be more specialized and specific than everyday language. When a prompt contains academic language, it often signals that the topic or subject matter is more specialized or technical in nature. The use of specific terminology and jargon in the prompt can help to define the scope and context of the question, making the generated responses more targeted and tailored to the given topic. By incorporating a higher degree of academic language use in prompts, the chatbot can produce more specialized and technical responses that align with the academic context. This can be especially important in educational settings where the topics discussed are often complex and specialized.

In summary, the results of the PCA analysis showed that the linguistic features related to LACS, CRD, specificity, and academic language use were the most prevalent features that characterized the corpus of efficient educational prompts that were known to produce high quality responses from ChatGPT.

### 3.7 Composite Score: Linguistic Prompt Efficiency Score (LPES)

Using the results of the PCA, a composite score is introduced in this study and derived from the PCA loadings and constituent indices. PCA loadings represent the relative importance of each component in explaining the variance in the data, while the constituent indices represent the specific linguistic features that contribute to each component. By multiplying the loadings with the normalized scores of the constituent indices for each prompt, a composite score that takes into account the relative importance of each component and the specific linguistic features that contribute to its efficiency is introduced in this paper.

**Composite score (LPES) = (LACS * 0.32) + (CRD * 0.16) + (Specificity * 0.12) + (Academic Language Use * 0.12)**

With this prompt composite score, a more quantitative metric is introduced in terms of how prompt efficiency can be approached through linguistic features. Prompt engineers will be able to tailor their lexical choices when crafting efficient prompts using the emergent components in this study and the LPES composite score. However, the researcher would like to highlight that the PCA done in this study only resulted to 4 components that represent up to 60% of the variance in the prompt corpus. Factors that can account for the remaining 40% need to be investigated further. To evaluate the perplexity scores of the generated responses for prompts with high LPES, the following pipeline was implemented (see figure 5).

**Figure 5. LPES evaluation pipeline**

A prompt with high LPES is shown below:

"I want you to act as a travel guide. I will write you my location and you will suggest a place to visit near my location. In some cases, I will also give you the type of places I will visit. You will also suggest me places of similar type that are close to my first location. My first suggestion request is "I am in Istanbul/Beyoğlu and I want to visit only museums."
The prompt is specific in its request for a place to visit near the location provided, with the added specification of museums only. The prompt is clear in its instructions, with a straightforward request and a clear indication of the type of place the user is interested in. The prompt also exhibits academic language use, as it is presented in a formal and technical manner, with precise language and specific terms such as "location," "place to visit," and "museums." These elements provide a clear and unambiguous description of what the user is looking for, which helps the chatbot generate a more effective and relevant response. Overall, the prompt is well-crafted and meets the requirements for an effective educational chatbot prompt. The generated response of the aforementioned prompt is presented below with a perplexity score of 34.49.

Sure, as a travel guide, I can suggest some museums for you to visit in Istanbul/Beyoğlu. There are several museums in the area that are worth visiting, including the Istanbul Modern Museum, the Pera Museum, and the Museum of Innocence. All of these museums have a unique collection of art and artifacts that showcase the rich cultural heritage of Istanbul. Additionally, if you’re interested in visiting more museums, there are several other options nearby that you might want to check out, such as the Istanbul Archaeological Museum, the Topkapi Palace Museum, and the Hagia Sophia Museum. These museums are all located within a short distance of Beyoğlu and offer a fascinating insight into the history and culture of Istanbul.

A vague prompt was written (shown below) because the corpus of prompts used in this study is comprised of known well-written and effective prompts. This ill-written prompt got a low LPES score and observably low presence of the constitutive elements is shown below:

“Act as a data scientist, design a hybrid recommender system approach for a real world use case.”

The generated response (excerpt shown below) was detailed and informative, but the perplexity score was 59.44, i.e. showing that ChatGPT has low confidence in its generated response.

As a data scientist, I can help design a hybrid recommender system approach for a real-world use case. Let’s consider a scenario where we are designing a recommender system for an online marketplace that sells a variety of products, ranging from electronics to clothing. To design a hybrid recommender system for this use case, we can combine collaborative filtering and content-based filtering. Collaborative filtering can be used to identify groups of similar users based on their purchase history, ratings, and other behaviors…

It was observed that the math prompt, even as it is well-written, had the lowest LPES and had a relatively high perplexity. Finally, a correlation analysis was performed on the sub-corpus of 200 randomly sampled prompts. The results indicate a significant negative correlation between LPES and Perplexity score, with a correlation coefficient of -0.65 (p < 0.05). This result indicates a meaningful relationship between LPES and Perplexity score. Additionally, the scatterplot of the data shows a clear downward trend, further supporting the negative correlation between LPES and Perplexity score. The LPES scores for the sample of 200 well written prompts ranged from -0.42 to 1.21 (meanLPES = 0.68; sdLPES= 0.04).

4. Conclusion
Understanding how to engineer prompts that are comprehensive and accurate can aid in the development of more effective educational materials. By leveraging these linguistic features in engineering prompts, learners and teachers can be more intentional in formulating prompts that leverage chatbots, large language models, and other generative AI tools. The four components of language in prompts reveal that efficient prompts are characterized by intentional grammatical structures (LACS and CRD), specificity, and academic language
use. Prompts with high values for both LACS (component #1) and CRD (component #2) are likely to be better grammatically structured and more natural prompts compared to those with low values for both measures.

Prompts with high LACS (component #1) indicate that the target word in the prompt has a strong tendency to occur with specific collocates, which means that the prompt is more likely to be grammatically well-formed such that the message is made clearer because of the grammatical structure. Prompts with high CRD (component #2) indicate that the words in the prompt tend to occur together frequently, which suggests that they are semantically related and likely to form a coherent sentence or idea. Conversely, prompts with low LACS and CRD values suggest that the words in the prompt are not closely associated with each other, and may not form a coherent or natural-sounding sentence. In this case, the generated prompt may be less useful or effective for its intended purpose. Overall, higher values for both LACS and CRD indicate that a chat prompt is more likely to be natural, coherent, and effective for generating meaningful and engaging conversations because of the grammatical decisions when formulating the prompt. Chatbots’ ability to generate better answers to prompts that are semantically coherent as a result of the well-formed grammar is due to the training on natural language patterns, also a grammatically well-formed knowledge base or training base, to generate related responses in the same context. Specificity (component 3) measures the degree of clarity of prompts through three constitutive indices: sentence length, unit length, and verb phrases. Efficient chatGPT prompts should be specific and clear to generate high-quality responses. Component 3 helps achieve this by measuring the clarity and specificity of the prompts through the three constitutive indices. By measuring sentence length, unit length, and verb phrases, the component evaluates how much information is packed into each sentence. In addition, using specific verbs can help enhance the clarity and specificity of the message being conveyed, and longer and more complex prompts may provide more context and specificity. However, the quality of a response ultimately depends on various factors beyond just the length, including the relevance and clarity of the prompt and the accuracy of the information provided. Overall, Component 3 contributes to efficient chatGPT prompts by providing a means to measure and improve the specificity and clarity of the prompts by incorporating just enough elaboration, hence having relatively longer prompts.

Lastly, academic language use is precise, specific, and technical, making it more specialized than everyday language. The presence of academic language in a prompt indicates a more technical or specialized topic. Using academic language helps define the context and scope of the question, enabling chatbots to generate targeted and tailored responses. This results in more accurate and precise answers. Component 4 is crucial in educational settings, where topics are complex and require specialized knowledge, producing more specialized and technical responses that align with academic context.

In conclusion, the linguistic attributes of efficient prompts that were identified in this study have important implications for learning. By leveraging these attributes, learners and teachers can be more intentional in their lexical choices when they leverage generative AI and similar tools in the teaching and learning process.

4.1 Limitations
The dataset used in this study was limited to publicly available corpora of known efficient prompts that teachers, researchers, and practitioners used. There was no comprehensive validation of the actual quality of the prompts aside from the perplexity estimate validation of its generated responses. Further the corpora used in this study do not represent the full range of possible prompts and topics that can be encountered in real-world applications and deployments of generative AI in education.

5. Future Work
There are several areas of future work that can build on the findings of this study. One potential avenue is to develop a tool or framework that can automatically identify high-quality prompts based on the linguistic features identified in this study. Such a tool could be used by educators and students to create more effective prompts that can lead to more accurate and comprehensive responses from LLMs. While the scope of this study is prompt engineering for chatbots, future directions in education can scale the benefits of the composite score (LPES) that is introduced in this study. LPES can be used to evaluate the linguistic efficiency of prompts used in educational chatbots or tutoring systems that are built on top of foundational models, helping to identify which prompts are most effective to facilitate optimal information retrieval.

References


Large Language Models (GPT) for automating feedback on programming assignments

Maciej PANKIEWICZ\textsuperscript{a*} & Ryan S. BAKER\textsuperscript{b}
\textsuperscript{a}Institute of Information Technology, Warsaw University of Life Sciences, Poland
\textsuperscript{b}Penn Center for Learning Analytics, University of Pennsylvania, USA
\textsuperscript{*}maciej_pankiewicz@sggw.edu.pl

Abstract: Addressing the challenge of generating personalized feedback for programming assignments is demanding due to several factors, like the complexity of code syntax or different ways to correctly solve a task. In this experimental study, we automated the process of feedback generation by employing OpenAI’s GPT-3.5 model to generate personalized hints for students solving programming assignments on an automated assessment platform. Students rated the usefulness of GPT-generated hints positively. The experimental group (with GPT hints enabled) relied less on the platform’s regular feedback but performed better in terms of percentage of successful submissions across consecutive attempts for tasks, where GPT hints were enabled. For tasks where the GPT feedback was made unavailable, the experimental group needed significantly less time to solve assignments. Furthermore, when GPT hints were unavailable, students in the experimental condition were initially less likely to solve the assignment correctly. This suggests potential over-reliance on GPT-generated feedback. However, students in the experimental condition were able to correct reasonably rapidly, reaching the same percentage correct after seven submission attempts. The availability of GPT hints did not significantly impact students’ affective state.

Keywords: Programming, automated assessment tools, automated feedback, LLM, GPT

1. Introduction

Large Language Models (LLMs) are deep neural network models able to effectively process and analyze complex linguistic structures (Carlini et al., 2021). LLMs have garnered significant attention in the field of natural language processing (NLP) because of the ability to generate human-like text. This feature makes LLMs a promising technology in educational settings (Finnie-Ansley et al., 2022; Dai et al., 2023; Pardos & Bhandari, 2023), where the provision of personalized feedback is integral to scaffold learning effectively (Jackson & Graesser, 2007; Hull & du Boulay, 2015).

One of the most prominent examples of LLMs is the OpenAI’s Generative Pre-trained Transformer (GPT) series (Radford et al. 2019; Brown et al. 2020), which leverages a transformer architecture and is able to capture long-range dependencies and positional information within a text. It is available through the ChatGPT application and via an Application Programming Interface (API), enabling integration with different existing applications, including educational apps. Especially in the field of computer science education, positive effects of such embedment may emerge earlier than in other educational domains, because of the significant effort conducted to develop models trained on large sets of programming code examples (Finnie-Ansley et al., 2022).
The field of computer science education has already extensively leveraged software tools that automate the process of self-paced learning. Multiple automated assessment tools have been designed to evaluate and provide feedback on student performance in various educational tasks (Deeva et al. 2021). The proliferation of these tools in computer science education, in areas from introductory programming (Edwards & Murali, 2017; Brusilovsky et al., 2018) to databases (Stanger, 2018), has had benefits and has contributed to the development of self-paced learning environments in other domains (Paiva et al., 2022). Along with this development, there is a growing body of research on personalizing feedback to enhance student learning outcomes (Deeva et al. 2021). Nevertheless, effective implementation of this approach continues to present significant challenges, given the difficulty of comprehensively assessing student work (Maier & Klotz, 2022).

In this study, we try to leverage the potential of GPT models for supporting students in providing feedback on programming tasks by integrating GPT through an API with an application for automated assessment of programming code. The GPT-3.5 model is used to automatically generate personalized feedback for university students taking an Object-oriented programming course. In this paper, we presented a controlled experiment comparing a system that adds GPT-generated hints to a system that only uses the original, human-developed hints. We compare the two conditions across multiple dimensions, including immediate performance, usage, performance on later content (without GPT-generated hints), time taken to complete assignments, and affect.

### 2. Methods

The study involved second-semester computer science students at the Warsaw University of Life Sciences (Poland) enrolled in an Object-oriented programming course that was required for their major. The course was conducted using the C# programming language. A total of 132 students consented to participate (out of 174 students taking the course) and were randomly assigned to either the control (N=66) or the experimental group (N=66). A pre-test was administered at the beginning of the semester to establish a baseline understanding of the students' knowledge before engaging with the platform's content. The process of data collection spans the initial part of the semester, specifically covering a nine-week period within the fifteen-week academic term. This timeframe extends from the first week of March 2023 through to the first week of May 2023. 93 students submitted at least one solution during the period of the study (control: 46, experimental: 47).

The experiment utilized the RunCode online application – a platform for automated execution and testing of a programming code available at runcodeapp.com (Pankiewicz, 2020). The application has been used by students at the University since 2017 within programming courses. All students that participated in this experiment were familiar with the application and had actively used it during the Introduction to programming course taught in the previous (first) semester. Students submitted programming code using an editor integrated within the application.

46 programming assignments covering basic object-oriented programming concepts (classes, objects, fields, methods, constructors, encapsulation, inheritance, and polymorphism) have been made available on the platform for the purpose of practicing the course material. These concepts were introduced during the first 6 weeks of the course. Each assignment was composed of a collection of more specific subtasks, such as: the creation of a class, the addition of fields with appropriate access modifiers, the definition of constructors, etc. A comprehensive suite of 809 unit tests was developed to thoroughly assess performance on these subtasks, ensuring a detailed evaluation of their individual components. Multiple submissions on a task were allowed. Usage of the platform was voluntary, and neither usage nor results within the platform counted towards the final grade. 5923 code submissions were collected during the period of study (control: 3077, experimental: 2846).

The unit tests were designed to validate code requirements specified by the assignment. For instance, a test aimed at confirming the presence of the "User" class within the submitted code might be denoted as `TestIsClassDefined("User")`, with an expected value of true. This
approach ensured that different aspects of the assignment's requirements were thoroughly
examined and verified in the students' code submissions.

The application provided students with a score (0-100%) and feedback after each code
submission. If the submitted code failed to compile, the regular feedback available on the
platform was presented without requiring any additional clicks and included details on compiler
errors. The platform also highlighted in the online code editor lines where errors occurred. If
the code compiled successfully, the feedback contained information on the executed unit tests,
including input values and expected output. Students needed to click on a specific test from
the list to access detailed information about its execution. The tests were color-coded, with
green indicating success and red signifying failure, to facilitate easy identification of the test
outcomes.

Both the control and experimental groups had access to the same set of tasks within the
application. However, for 38 out of 46 tasks, students in the experimental group had an
additional feature enabled, which provided them with feedback generated by a large language
model through the GPT-3.5 API. For the remaining 8 tasks, students in the experimental group
only had access to the regular feedback offered by the platform. These tasks were slightly
more challenging and were designed to encapsulate the concepts introduced in previously
solved tasks, as well as assess the comprehension of these concepts.

Within the experimental condition, additional feedback was provided by the GPT model,
going beyond the standard feedback provided by the application, which included information
on compiler errors, runtime errors and details on executed unit tests. The GPT-generated
feedback provided suggestions for code improvement, explanations of compiler errors, and
hints for debugging. The feedback was designed to be informative and constructive, aiming to
guide students towards correct solutions and improve their understanding of programming
concepts, without revealing the correct code solution. This feedback was presented
immediately after code submission, with a prompt asking students to rate the usefulness of
the hint on a 5-point Likert scale ranging from 'Not useful at all' to 'Extremely useful'. Feedback
was automatically generated in Polish, to match the assignments and the interface of the
system, which also used this language.

Feedback was requested from GPT when the code submitted by a student exhibited
compiler errors, runtime errors, or failed at least one unit test. The dynamically generated
prompt included the assignment text (in Polish) and the student's code, accompanied by
additional information based on the testing results. This information contained a compiler
message for non-compiling code, exception type and message for compiled code that
generated an exception during execution, or details regarding a failed unit test: the test name,
input values, and expected outcome. To accommodate the OpenAI GPT API's token limitation
(4,000 tokens), assignments were designed in a way that ensured that even the most complex
tasks and accompanying code could fit within the imposed constraint. The prompt text was
similar for the three scenarios mentioned, with variations stemming from the testing process
outcomes. In cases where the code did not compile, a specific prompt was generated:

I want you to act as a Stackoverflow post that helps me to solve a programming
assignment in C#. I want you to explain in Polish why this code does not compile.
Don't write solution in the explanation, but focus on meaningful hints. I want you
to also include the line where the compiler error occurred in the explanation. I
want you to also include a line number for each detected error in the explanation.
To help me better understand your response, highlight keywords, line numbers, class
names, variable names, messages, line numbers and error names with the <code> </code>
HTML markup in the explanation. Programming assignment: ### <ASSIGNMENT_TEXT> ###
C# code: <STUDENT_CODE> ### Compiler errors: <COMPILER_ERRORS>

The following parameters were used for the request: 'model': 'text-davinci-003',
'temperature': 0, 'max_tokens': 500, 'n': 1, 'top_p': 1, 'frequency_penalty': 0,
'presence_penalty': 0.

An illustration of the hint proposed by the GPT in the event of encountering code with a
syntax error is presented as follows (translated from Polish to English with highlights
generated by the GPT): “A compiler error error CS1002 occurred in line 5 of the code with
the message ; expected . This means that there is a missing semicolon ; after the
declaration of the `string name` field in line 5. To fix the error, a semicolon `;` should be added at the end of line 5. A `;` is required at the end of every declaration in C#.

Students self-reported their affective states while completing tasks on the platform by interacting with a dynamic HTML element. The platform prompted students for their emotional state following the receipt of submission results by asking them to select the option that best described their current feelings. The response options included: Focused, Anxious, Bored, Confused, Frustrated, and Other (in this order), accompanied by representative emoticons. These affective states were chosen due to their relevance to the learning process and their prevalence in past research (Karumbaiah et al., 2022). To mitigate potential frustration from overly frequent survey prompts, the platform randomly decided after each submission whether to present the survey, with a probability of one in three.

3. Results

Due to violations of normality assumptions, non-parametric Mann-Whitney U tests were utilized. Benjamini-Hochberg alpha correction (Benjamini & Hochberg, 1995) was applied to control for multiple comparisons in analysis of differences in affective state frequencies among students.

3.1 Pre-test results

To ensure that the control and experimental groups were comparable in terms of their initial understanding of the relevant programming concepts at the beginning of the study, a pretest consisting of 7 multiple-choice questions was administered to both the control and experimental groups. These questions covered key topics on object-oriented programming, including classes, constructors, encapsulation, inheritance, and polymorphism. No statistically significant differences were identified between the control (Mdn=0) and the experimental (Mdn=1) groups (W=1791, p=0.249, Mann-Whitney U test).

3.2 Usefulness of the GPT hints

In the examination of the perceived usefulness of the GPT generated hints, a total of 1,442 responses were collected from N=46 participants within the experimental group throughout the duration of the study. As the control group did not have access to the GPT hints, no responses were obtained from this group. Figure 1 displays the histogram of the frequency of users according to the median rating they gave to received hints, with the x-axis representing the median of the hint rating for a user.

![Figure 1. Distribution of users according to the median rating for received hints.](image-url)
As Figure 1 shows, the majority of users rated the GPT hints they received as useful – these results suggest that the hint feature was generally well-received by users.

### 3.3 Usage of the regular platform feedback

One possible impact of the GPT feedback is that learners may use the other feedback less often. To evaluate whether there were differences in the usage of the regular (non-GPT) platform feedback between the experimental and control groups, we analyzed the percentage of incorrect submissions for which students clicked on feedback to access details about tests that failed. We did not examine clicks on feedback for tests that ended successfully, as the application did not generate GPT hints for these tests. Therefore, students who only made successful submissions, and students who submitted only non-compiling code in their submissions (in this case the platform always presents feedback, so users are not required to click on it), were not included in this analysis.

A statistically significant difference was found between the experimental (Mdn=0.321, N=45) and control (Mdn=0.710, N=45) groups (W=282, p<0.001, Mann-Whitney U test) for the set of tasks where experimental group had GPT feedback enabled (38 out of 46). The experimental group used the regular feedback significantly less than the control group for these tasks.

However, no significant difference was found between the experimental (Mdn=0.769, N=23) and control (Mdn=0.667, N=25) groups (W=341, p=0.269, Mann-Whitney U test) for the set of tasks, where GPT feedback was not available for experimental group (8 out of 46). When GPT feedback was unavailable, the experimental group utilized the platform’s regular feedback in a similar amount to the control group (Figure 2).

![Figure 2](image)

*Figure 2. Comparison of the percentage of incorrect submissions with at least one click to request (view) details on the test that failed (regular platform’s feedback) between the control and experimental group (on the left: tasks with GPT feedback enabled for experimental group; on the right: tasks with GPT feedback disabled for the experimental group).*

### 3.4 Performance within the platform – GPT hints enabled

In order to examine the influence of the GPT hints on student performance within the platform, we conducted an analysis of users’ consecutive task attempts for tasks where GPT hints were enabled for the experimental group (38 out of 46 tasks). The primary focus was on assessing the percentage of successful submissions after each attempt (cumulative).

For the first attempt, the experimental group appeared to have a similar percentage of successes (48.3%) as the control group (48.0%). Values for the second and all consecutive attempts decreased for both groups, with performance worsening across attempts. Following the second attempt, the experimental group appeared to have a slightly higher percentage of
successful submissions, a trend that persisted across subsequent attempts. For the purpose of clarity, only the first 15 attempts are depicted in the chart (Figure 3).

Figure 3. Comparison of the percentage of successful submissions at each attempt (cumulative) for the control and experimental group, for tasks where GPT hints were enabled in the experimental group.

We employed a linear mixed-effects model to test whether there was a significant difference in slopes between the control and experimental groups concerning the number of attempts. The model was fit using restricted maximum likelihood (REML) estimation with the \texttt{nlme} R package (Pinheiro et al., 2017). The fixed effects of the model consisted of the main effects of group and attempt, as well as the interaction between group and attempt. Our analysis revealed a non-significant interaction effect between group and attempt, $t(13)=-0.172$, $p=0.866$. This indicates that the slopes of the control and experimental groups do not significantly differ. Nonetheless, the main effects of group and attempt were found to be statistically significant, group: $t(13)=5.38$, $p<0.001$; attempt: $t(13)=-7.38$, $p<0.001$. This finding suggests that, averaged across all attempts, the experimental group scored higher than the control group on the dependent variable.

3.5 Performance within the platform – without GPT hints

To further evaluate the impact of the introduced platform feature on the experimental group, we analyzed performance of both groups on the tasks where GPT-generated hints were not provided for the experimental group (8 out of the total 46 tasks).

As in the prior section, we visualized the percentage of successful submissions following each attempt. For the initial six attempts, the control group performed better than the experimental group; however, starting from the seventh attempt correctness rates were comparable (Figure 4).

A linear mixed-effects model fit using restricted maximum likelihood (REML) was employed to assess whether the slopes of the control and experimental groups were significantly different with respect to the number of attempts. The model revealed a significant
interaction effect between group and attempt, \( t(13)=7.13, p<0.001 \). This finding indicates that the slopes of the control and experimental groups were significantly different.

\[\text{Figure 4. Comparison of the percentage of successful submissions at each attempt (cumulative) for the control and experimental group for tasks where GPT hints were disabled in the experimental group.}\]

3.6 Time needed to successfully complete tasks

In order to assess the differences in the time required for students in the experimental and control groups to successfully complete tasks on the platform, we analyzed 1,784 successful student submissions during the study period (control: 905, experimental: 879).

We evaluated the time taken by students in both conditions to fully solve each task, while excluding tasks completed by fewer than three students in each condition. Due to the platform’s lack of keystroke-level data monitoring, a time limit of 7,200 seconds (2 hours) was established for tasks not solved within this duration. This decision was made because some students ceased working on tasks without logging out.

A significant majority (88%) of successful student attempts were completed within this time frame. Due to violations of normality assumptions, we used rank-based regression—a non-parametric alternative to traditional likelihood or least squares estimators (Kloke & McKean, 2012).

We analyzed the time performance on tasks where the GPT feedback was available for the experimental group. For this set of tasks, the group was not a statistically significant predictor of the time needed to solve tasks, \( t(89)=-0.24, p=0.811 \). However, the pretest score was a marginally significant predictor, \( t(89)=-0.69, p=0.095 \).

We also conducted the analysis for tasks where the GPT feedback was not available for the experimental group. The group was a statistically significant predictor of the time needed to solve tasks, with the experimental group needing less time than the control group, \( t(48)=-2.25, p=0.029 \). The pretest score was not significantly associated with time needed to solve tasks, \( t(48)=-0.36, p=0.723 \). For the set of these tasks results indicate that students in the experimental group took, on average, 375 seconds (6.25 min.) less time than the control group, controlling for the pretest score.
3.7 Affective states

A total of 1,304 affect survey responses were collected from participants over the course of the study (645 in the experimental condition, 659 control). In the analysis, we include students who submitted at least 3 responses in this survey: in the control N=33, and in the experimental group: N=35.

Both conditions predominantly reported the focused state, which constituted over half of the responses in each group. No significant differences between conditions were observed for the reported frequency of any of the affective states (Table 1).

Table 1. Affective State Survey – Statistical Evaluation of Differences in Frequencies of Affective States Reported by Students in Experimental and Control Group

<table>
<thead>
<tr>
<th></th>
<th>Experimental (Mdn/Mean)</th>
<th>Control (Mdn/Mean)</th>
<th>W statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused</td>
<td>0.538/0.52</td>
<td>0.5/0.54</td>
<td>537</td>
<td>0.620</td>
</tr>
<tr>
<td>Frustrated</td>
<td>0/0.14</td>
<td>0.13/0.05</td>
<td>611</td>
<td>0.653</td>
</tr>
<tr>
<td>Anxious</td>
<td>0/0.12</td>
<td>0/0.12</td>
<td>550.5</td>
<td>0.718</td>
</tr>
<tr>
<td>Confused</td>
<td>0/0.06</td>
<td>0/0.09</td>
<td>571</td>
<td>0.935</td>
</tr>
<tr>
<td>Bored</td>
<td>0/0.03</td>
<td>0/0.05</td>
<td>559.5</td>
<td>0.782</td>
</tr>
</tbody>
</table>

Frustration did not significantly differ between the control (Mdn=0) and experimental (Mdn=0) groups (W=611, adjusted α=0.02, p=0.653, Mann-Whitney U test).

Similarly, no significant difference was found for boredom between the experimental (Mdn=0) and control (Mdn=0) groups (W=550.5, adjusted α=0.04, p=0.782, Mann-Whitney U test).

The same pattern was observed for anxiety and confusion, with no significant difference between the experimental (Mdn=0) and control (Mdn=0) groups for anxiety (W=550.5, adjusted α=0.03, p=0.718, Mann-Whitney U test) or confusion (W=571, adjusted α=0.05, p=0.935, Mann-Whitney U test).

Focused also exhibited no significant differences between conditions, with the experimental group's median (Mdn=0.538) not statistically different from the control group (Mdn=0.5; W=537, adjusted α=0.01, p=0.620, Mann-Whitney U test).

4. Discussion and summary

In this study, we utilized the GPT-3.5 model to generate personalized hints for students working on programming assignments within an automated assessment platform. Our findings indicated that almost half of the students (46%) highly valued the usefulness of GPT-generated hints, with a median rating of 4 or 5 on a 5-point Likert scale. Given that the assignments and generated hints were provided in Polish, these findings hold promise for the future development and scalability of such systems across various languages, further extending their applicability and impact in diverse educational contexts.

However, 19% of the students found the hints to be not useful, with median ratings of 1 or 2 on the same scale. Thus, there is still room for improvement and for making the hints more useful. Considering the varied ratings of hint usefulness, we anticipate that further optimization of the hint generation process could enhance the efficacy of GPT-based automated feedback, ultimately leading to improved student performance. Future research may be able to further optimize and enhance the hints, to increase the proportion of useful hints and the degree to which each hint is useful. Potential strategies for achieving this include more experimentation with prompt generation (prompt engineering), fine-tuning the GPT model (by providing examples on which the model can improve) or conducting correlation mining to identify the properties associated with less useful hints.

Another finding is that the experimental group (with GPT hints enabled) relied less on the platform's regular feedback for tasks where the GPT feedback was enabled. Despite of
significantly lower usage of this kind of feedback, the experimental condition students performed significantly better on the platform in terms of percentage of successful submissions across consecutive attempts. There was not a significant difference in the time needed to successfully solve each assignment with GPT feedback enabled. However, students in the experimental group spent an additional average of 40 seconds after each attempt compared to the control group. This extra time was used by the students to read and understand the hint that was generated before they proceeded to rate it through a survey.

Furthermore, when GPT hints were made unavailable, students in the experimental group needed significantly less time to successfully solve each assignment. Since hints enabled for earlier tasks not only served to rectify the students' misconceptions but also provided supplementary information, it is our conjecture that these hints played a pivotal role in enhancing the students' learning process, apparently further facilitating their understanding. This knowledge acquisition, in turn, may have expedited their task-solving abilities, leading to the observed reduction in time required to complete tasks when the hints were no longer available.

An unexpected outcome of this study is that when students in the experimental condition switched over to more difficult tasks for which GPT hints were not enabled, these students were significantly less likely to successfully solve the tasks in their first several attempts. Our findings indicate that users appeared to rely on the GPT-generated feedback during the study and needed some time to adapt to a lack of this kind of feedback. This may indicate that students were becoming overly dependent on the GPT-generated feedback, but if so, this problem seemed to correct itself fairly rapidly.

Although the GPT hints impacted student performance, the availability of these hints did not appear to have a substantial impact on student affect.

This study presents several limitations that warrant further investigation. First, the study duration spanned only 9 weeks of the 15-week semester term; future research should explore the longer-term effects of GPT-generated hints on student performance and learning outcomes. Second, the experiment focused solely on GPT-generated hints in the Polish language, raising the need for additional experiments in different languages to establish the generalizability of the results. Third, the number of participants represents a limitation to the robustness and external validity of our findings; future work should examine this type of intervention with a larger sample size. Finally, the study utilized a single programming language (C#), while introductory computer science education often incorporates other languages, such as Java and C++. Future research should expand the scope to include a broader range of programming languages to more comprehensively assess the efficacy and applicability of GPT-generated hints.

Addressing the challenge of generating personalized feedback for programming assignments is a demanding task, given the myriad of potential code syntax errors and code complexity across different assignments. The GPT model holds the potential to tackle this issue more effectively and efficiently than human authoring of hints. In summary, the findings of this study indicate that integrating GPT-generated feedback into computer programming education may positively impact student performance, ultimately contributing to enhanced learning outcomes.

Although this experiment employed the default GPT-3.5 model, it is plausible that fine-tuning the model could further improve the quality of generated hints, thereby facilitating more effective programming education on a larger scale.
Acknowledgements

This paper was written with the assistance of ChatGPT, which was used to improve the writing clarity and grammar of first drafts written by humans. All outputs were reviewed and modified by two human authors prior to submission.

References


Improved Automated Labeling of Mathematical Exercises in Japanese

Taisei YAMAUCHI\textsuperscript{a}, Ryosuke NAKAMOTO\textsuperscript{b}, Yiling DAI\textsuperscript{b}, Kyosuke TAKAMI\textsuperscript{b,c}, Brendan Flanagan\textsuperscript{d}, & Hiroaki OGATA\textsuperscript{b,c}

\textsuperscript{a}Graduate School of Informatics, Kyoto University, Japan
\textsuperscript{b}Academic Center for Computing and Media Studies, Kyoto University, Japan
\textsuperscript{c}Education Data Science Center, National Institute for Educational Policy Research (NIER), Japan
\textsuperscript{d}Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Japan

*Yamauchi.taisei.28w@st.kyoto-u.ac.jp

Abstract: This study aims at improving the prediction quality of the automatic labeling of learning materials. Labeling learning materials has two existing issues: establishing completely automated labeling and reducing manual labor for assigning labels to materials. Labels of the materials are utilized for analyzing students’ learning patterns, tracing knowledge, and recommending exercises to students. Since it is too burdensome to manually assign several labels to many learning materials, an automatic, algorithm-based labeling system is desirable. However, classification using word embedding has often yielded lower accuracy for mathematics learning materials with short texts. In this research, we have conceived and implemented an improved approach to predict a label by calculating the similarity of n-gram of sentences using Jaccard coefficients, weighting them to create a vector representation, and using it to predict the label of the exercise. We compared the accuracy and F score of the prediction results of the weighted n-gram similarity model with those of the state-of-the-art word embedding model. We found that the n-gram approach was superior in both accuracy and F score. Furthermore, we plotted the vectors obtained from each model in two-dimensional coordinates and observed that the n-gram model produced more flexible predictions, regardless of the vector’s position. These results suggest the classification effectiveness with weighted similarity of n-gram for materials with a small amount of text.

Keywords: Automatic labeling, word n-gram, Jaccard coefficient, t-SNE, word embedding, topic-based learning.

1. Introduction

Labeling learning materials is a widely used method in educational domains, such as knowledge tracing to adapt how to teach to each student (Vie & Kashima, 2019), automatic exercise recommendations to support students’ learning strategy (Takami et al., 2022) and analysis based on the material topic (Wang et al., 2022). With the shift to ICT education, to predict how well the students tackle exercises (Vie & Kashima, 2019), to recommend the most appropriate exercise in many learning materials (Takami et al., 2022) or to utilize them for learning pattern analysis (Wang et al., 2022), a topic-based classification of instructional materials is critical to utilize them conveniently. However, assigning the classification to problems manually is a hard task that requires the cooperation of experts. In other words, it is desirable to label exercises automatically to reduce the cost and time of domain experts who usually perform the task. Previous classification studies have taken one reliable method: word embedding. A previous study has shown that word embedding has produced novel results for classifying news stories (Dharma et al., 2022). However, another previous research into mathematical exercise classification results in lower accuracy for exercises, each consisting of a small number of words (Tian et al., 2022).
In this study, we propose an automatic classification algorithm that can handle even short sentences in mathematics exercises. We focused on the optimal agreement of a set of mathematical problems and calculated the similarity between the weighted word n-gram variance representation of any labeled exercise and that of each exercise query. For further understanding of the results, we use t-SNE (Van der Maaten & Hinton, 2008) and compared the proposed bi-gram vector space model with that of novel word embedding. We set the following research question and tackle it:

**RQ:** Can the use of a weighted n-gram classifier improve the labeling of Japanese mathematics quizzes compared to previous research using state-of-the-art word embedding methods?

## 2. Literature Review

There is a trend toward analyzing learning behavior in a new way using the labels assigned to teaching materials. With regard to the use of features in learning effectiveness analysis, a study reported that the proposed system automatically assigned labels with learning materials and analyzed the assigned labels to discover students' learning patterns (Wang et al., 2022). Giving metadata to exercises for knowledge tracing is also a hot research topic. One study, using multiple real data sets consisting of tens of thousands of users and items, showed that regression classification models could accurately and rapidly estimate student knowledge, even when student data is sparsely observed. In addition, the study showed that the model can handle multiple knowledge elements and side information such as the number of trials of items and skill levels (Vie & Kashima, 2019). It is also useful to categorize many exercises for recommending a specific exercise to enhance students’ understanding. One study developed a recommendation system with explanations based on an explanation generator using parameters from a Bayesian knowledge tracing model (Takami et al., 2021). The recommendation quiz with explanations was used more frequently and with more continuous users compared to a system without explanations (Takami et al., 2022). There has also been researching into extracting labels from learning materials to form knowledge structure representations that can be shown to learners to increase their awareness of the study process (Flanagan et al., 2019). These research examples show the importance of providing labels to materials because it is easier to obtain or utilize detailed information about the characteristics of the material if they are labeled in advance.

Studies on automated labeling or classifying methods also focus on reducing the burden of manual labeling. A study has attempted to classify using 385 different labels to classify 12 years of mathematics materials from kindergarten through high school (Shen et al., 2021b), whose classification categories are according to the common core standard developed by the United States (Ritter, 2009). The label of the study is for K-12 Education in the United States, which is not somewhat appropriate for the Japanese one. One study proposed the MathBERT model, a model created by pre-training the base BERT model (Devlin et al., 2018) on a large corpus of mathematics content ranging from kindergarten students (pre-k) to high school and university graduate-level mathematical content (Shen et al., 2021a). Labeling and classifying instructional materials automatically is receiving much attention, and several researchers have investigated methods to solve this important task. One paper proposed an automatic classification method in a mathematical subject classification scheme for organizing mathematical literature, achieving a classification accuracy of 81%, which is very close to the classification accuracy of two extensive peer-review services. It also enabled an 86% labor reduction compared to manual classification (Schubotz et al., 2020). This study shows the importance of labeling materials automatically to reduce the human labor of manual labeling.

Previous research has found that word embedding can be effectively applied to label classification tasks. Dharma, et al. (Dharma et al., 2022) found that word embedding using the Fasttext method was able to classify a dataset containing 19,977 news stories and 20 news topics with 97.2% accuracy, which is more accurate than any other word embedding method.
However, when it comes to classifying exercises with short sentences, sentence vectorization using word embedding has been found to be a less effective method. Tian et al. (Tian et al., 2022) applied word embedding as a method to classify short Japanese exercise texts and found it achieved 72.87% accuracy. Accuracy was then further improved by combining the method in an ensemble with a keywords extraction method. This suggests that the word embedding method might not be as effective for short exercise texts.

Therefore, in this study, we propose an automatic classification algorithm that can handle even short Japanese sentences in mathematics exercises by focusing on the optimal agreement of a set of mathematical problems by calculating the similarity between the weighted word n-gram variance representation of any labeled exercise text and the query exercise texts.

### 3. Method

While word embedding has been effectively applied to many tasks, labeling mathematics exercises in Japanese that contain few words is challenging and has not performed as anticipated. In this paper, in addition to classification by word embedding, we examine the classification by Jaccard coefficient word n-gram to see if it can provide improved performance in the task. For this purpose, we assigned an appropriate unit label to the exercises, trained them using an algorithm, and made predictions.

![Figure 1. Overview of the method.](image)

The details of the two approaches will be explained in the following sections as indicated in Figure 1. Note that we explain the details of the two approaches in the following sections.

### 3.1 Data Preprocessing

The exercise texts used in the experiment are from six different digital mathematical exercise books. These exercise books are for high school students in accordance with the textbooks designated by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2021), and are made by the same company that produced the textbooks, which are all well-used in Japanese. We prepared text files by reading text data from pdf using the Python library PyPDF2 (Palmer, 2021). Due to possible problems from file structure and text embedding methods, Pdf files are more difficult to obtain in their complete text form than HTML-formatted files, as previous studies confirm (Ramakrishnan et al., 2012; Smith, 2007).

We used a package called Nagisa (Ikeda, 2021) for the morphological analysis of text data. Nagisa is a package for the morphological analysis of Japanese sentences. This process is required to extract individual words because there are no explicit word separators such as blank spaces in Japanese. When Nagisa morphologically analyzes number digits, it recognizes a set of digits that each recognized as a noun. For treating different numbers as the same, we processed a set of digits into a single number and treated all numbers as the same thing.

Each exercise has the subject labels whose publisher allocated on, so we treated them as the correct labels. There was a total of 2,775 exercises, consisting of 24 subject labels, and each subject label was assigned between 25 and 200 exercises. Figure 2 shows an example of an exercise pdf data and the assigned subject label. The figure also shows a pair of exercise
text data with very similar sentences, and other pairs of problem data with similar sentences were found in the data set.

There are 24 subjects, with each having an average of 115.6 exercises. Each exercise contains M = 98.5, and SD = 61.4 morphemes, with a minimum of 17 and a maximum of 469 occurring in the dataset that was collected. This shows the limited amount of information contained in the dataset which the classifier can use to automatically assign labels. Compared to English, mathematics quizzes the number of morphemes is smaller.

![Figure 2](image)

Figure 2. Example of exercises in the dataset (Right: English translation). The subject label “number and formula” is assigned to an exercise in the figure.

3.2 Existing Word-embedding Prediction Method

For vectorization with word embedding, we used a model called fastText (Joulin et al., 2016). Among the fastText models for 157 languages, we used the Japanese model in this experiment. This model combines three methods to represent a single sentence data in 300 dimensions: character 5-gram, weighting by position, and Word2Vec (Church, 2017). By considering each word as a set of its constituent sub-words, it offers notably improved performance when dealing with languages that have complex morphological structures including Japanese (Khan et al., 2022). To evaluate classification performance using word embedding, we used 5 baseline models which have different characters. Five models we used are as follows:

- **XGBoost** (Chen & Guestrin, 2016): a model that combines boosting and decision trees and has shown good results in various natural language processing tasks, so it’s proper to use in the context of this paper.

- **Random Forest** (Breiman, 2001): a model that uses many decision trees trained by randomly sampled training data. This works well even with a large number of explanatory variables, so it can deal with a 300-dimensional vector.

- **Support Vector Machine** (Vapnik and Lerner, 1963): a model for constructing a two-or-more-class pattern discriminator using linear input elements. This is one of the most popular methods to classify something, so it is appropriate to use as a baseline.

- **Logistic Regression** (Cox, 1958): a statistical regression model with variables that follow a Bernoulli distribution. As well as SVM, it is one of the most popular methods to classify something, so it is appropriate to be treated as a baseline.

- **Multilayer Perceptron** (Gardner & Dorling, 1998): a model that uses feedforward artificial neural networks. A neural network is also treated as the baseline for the machine learning tasks, so we use it in this experiment.
3.3 Proposed N-gram Prediction Method

We created the word n-gram from the tokenized exercise sentences. The prepared data set is divided into label data and query (query means an unlabeled exercise), and the similarity between the set of word n-grams in the labeled data and the set of word n-grams in the query data is calculated. We described how to choose one subject for each not-yet-labeled exercise in Figure 3. In the figure, \( n_{t_{A_k}} \) s in green boxes represent the n-gram set of the labeled exercise text, and \( n_{t_q} \) s in blue boxes do that of a query. We evaluated the similarity \( J_{l_{x,k,q}} \) between \( n_{t_{A_k}} \) and \( n_{t_q} \) by calculating the Jaccard coefficients of those. Then, we weighted \( J_{l_{x,k,q}} \) s and found a value \( J_{l,q} \) corresponding to label \( l \). If the value \( J_{l,q} \) is the highest of all \( l \) of \( J_{l,q} \), then the label \( l \) was predicted.

For vectorization using the word n-gram, we used a function in our algorithm. As previous studies improve accuracy by weighting for realistic non-homogeneous data sets (Graovac et al., 2015; Kobayashi, 2021), we modified the weights of the calculated n-gram similarity to create a more accurate classifier. We created two functions based on two assumptions as follows:

- **Assumption 1:** Two exercises that have the same label are similar to each other. Therefore, we created \( f_{\text{mean}} \) (1) to find the most proper label considering all labeled exercises' similarity.
- **Assumption 2:** \( m \) specific exercises with the same label have high similarity with each other. Therefore, we created \( f_{\text{rank}_m} \) (2) to find the most proper label considering \( m \) labeled exercises' similarity.

We substitute \( J_{l_{k,q}} \) for all labeled data \( l_k \) with label \( l \) all into the determined function \( f \) to obtain the weight coefficient \( J_{l,q} \) of queries \( q \) that are assigned that label. Hereafter \( \text{LARGE}(J_{l,q}, k) \) represents the \( k \)th highest value of \( J_{l_1,q}, J_{l_2,q}, \ldots, J_{l_m,q} \). In this experiment, the following functions were defined to determine a more suitable weighting for classification:

\[
f_{\text{mean}} = \frac{\sum_{k=1}^{n} J_{l_{k,q}}}{n} \tag{1}
\]

\[
f_{\text{rank}_m} = \frac{\sum_{k=1}^{m} \left( (m-k+1) \times \text{LARGE}(J_{l,q}, k) \right)}{\sum_{k=1}^{m} k} \tag{2}
\]

We then predict a label \( l_{\text{pred},q} \) for each query exercise \( q \) by choosing the highest coefficient of \( J_{l,q} \):

\[
l_{\text{pred},q} = \arg \max_{l} J_{l,q} \tag{3}
\]

![Figure 3. The way to find the weighted coefficient and predict a label.](image)
3.4 Evaluation

To prevent bias in the dataset due to random sampling, we performed classification training with 80% of the entire dataset as training data and the rest as test data using 5-fold, a form of cross-validation. This allows us to train five times on one dataset.

To evaluate the effectiveness of our proposed method, we compare it to existing word embedding methods using the following two metrics: accuracy and F score. Accuracy and F scores on binary predictions are typically calculated by accurate prediction divided by all predictions and the harmonized average of precision and recall, respectively, but since the model in this study proposed here performs multi-class classification, it is necessary to use a calculation method customized for multi-class classification. For evaluation of the result in this study, the accuracy $A_L$ and the precision $P_L$, the recall $R_L$, and the F score $F_L$ for all labels were calculated using the following equation (Sorower, 2010).

$$A_L = \frac{\sum_{l \in L} A_l}{n(l)}$$
$$P_L = \frac{\sum_{l \in L} P_l}{n(l)}$$
$$R_L = \frac{\sum_{l \in L} R_l}{n(l)}$$
$$F_L = \frac{2P_lR_l}{P_l+R_l}$$

For further examine the experimental results, we also mapped the assigned labels to a two-dimension space for visualization by using t-SNE. The technique t-SNE is one novel technique that can project high-dimensional data into two dimensions and map them (Van der Maaten & Hinton, 2008). One study compared how well the clustering of the different methods with observing the visualization of the result with t-SNE, and found that one model that uses sentence embedding with knowledge graph embedding outperforms another model that only uses sentence embedding (Chen et al., 2022).

4. Result

4.1 Prediction Result

Figure 4 shows a plot of the accuracy for each weight function $f$ for each value of $n$ in the n-gram. We understand from the figure that the best accuracy was achieved for bigram ($n = 2$) for all functions $f$. Table 1 shows the accuracy and F score for each weight function for bigram. We found that the proposed algorithm predicts the best when we use the function $f_{rank7}$.

<table>
<thead>
<tr>
<th>Function $f$</th>
<th>Accuracy</th>
<th>F score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{mean}$</td>
<td>.7222</td>
<td>.7322</td>
</tr>
<tr>
<td>$f_{rank1}$</td>
<td>.8061</td>
<td>.8023</td>
</tr>
<tr>
<td>$f_{rank2}$</td>
<td>.8209</td>
<td>.8154</td>
</tr>
<tr>
<td>$f_{rank3}$</td>
<td>.8274</td>
<td>.8224</td>
</tr>
<tr>
<td>$f_{rank4}$</td>
<td>.8324</td>
<td>.8288</td>
</tr>
<tr>
<td>$f_{rank5}$</td>
<td>.8382</td>
<td>.8346</td>
</tr>
<tr>
<td>$f_{rank6}$</td>
<td>.8447</td>
<td>.8408</td>
</tr>
<tr>
<td>$f_{rank7}$</td>
<td><strong>.8483</strong></td>
<td><strong>.8441</strong></td>
</tr>
<tr>
<td>$f_{rank8}$</td>
<td>.8458</td>
<td>.8420</td>
</tr>
<tr>
<td>$f_{rank9}$</td>
<td>.8461</td>
<td>.8427</td>
</tr>
<tr>
<td>$f_{rank10}$</td>
<td>.8461</td>
<td>.8432</td>
</tr>
</tbody>
</table>
Table 2 shows the accuracy and F scores from each of these experiments using a word embedding and machine learning modules. We added the algorithm with the best classification accuracy in each n-gram experiment for comparison.

### Table 2. Accuracy and F score of the prediction using a word embedding with two essential results of the prediction using n-gram.

<table>
<thead>
<tr>
<th>Features</th>
<th>Models</th>
<th>Accuracy</th>
<th>F score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embedding</td>
<td>XGBoost</td>
<td>.6505</td>
<td>.6373</td>
</tr>
<tr>
<td>Word embedding</td>
<td>Logistic Regression</td>
<td>.6515</td>
<td>.6403</td>
</tr>
<tr>
<td>Word embedding</td>
<td>Perceptron</td>
<td>.6840</td>
<td>.6949</td>
</tr>
<tr>
<td>Word embedding</td>
<td>Random Forest</td>
<td>.7023</td>
<td>.6974</td>
</tr>
<tr>
<td>Word embedding</td>
<td>SVM</td>
<td>.7939</td>
<td>.7992</td>
</tr>
<tr>
<td>N-gram (n = 2)</td>
<td>Jaccard Coefficient + $f_{mean}$</td>
<td>.7222</td>
<td>.7322</td>
</tr>
<tr>
<td>N-gram (n = 2)</td>
<td>Jaccard Coefficient + $f_{rank7}$</td>
<td><strong>.8483</strong></td>
<td><strong>.8441</strong></td>
</tr>
</tbody>
</table>

### 4.2 Mapping Vector Representations with t-SNE

Figure 5 shows the mapping of the results of classification using bi-gram (Jaccard + $f_{rank7}$) and that using word embedding (XGBoost). In the figure, exercises with different labels are represented by different colors and symbols. In (a)I, (b)II of the figure, exercises with wrong predictions are mapped in gray color.

Looking at (a)I and (b)I in Figure 5, we can see that both have a group of exercises with the same labels in close proximity to some extent. We can observe that the center of (a)I have a very large number of exercise data with different labels distributed tightly together, and from (a)II that they are almost always incorrectly predicted. However, (b)I do not show such portions of the data. This observation suggests that the word embedding method struggles to accurately label quizzes that are close to many different types of exercises when compared to using the proposed n-gram classification method.

Looking at (a)II and (b)II in Figure 5, we see that for both sets of data where the predictions are correct, exercises with the same label are distributed in close proximity to each other to some extent. In (a)II, there are many areas where items of the same type are clustered together and a few areas where items of different types are close to each other. On the other hand, it can be observed that in (b)II, there are many locations where items of different types are located close to each other. For example to compare with the label distribution in the transparent square of a(II) and b(II), for labels with labels a, c, d, k, o, and u, in (a)II, each label appears to distribute in a definite region, and each region appears to have no correct predictions for data with other labels. In contrast, in (b)II, we often find small regions where the labels are different but the predictions are correct. This accounts for the fact that prediction by word embedding is sensitive to the position corresponding to the vector representation of
the data, whereas n-gram similarity is able to make predictions regardless of the position corresponding to the vector representation of the data.

Figure 5. Maps of the vectors obtained from each experiment were transformed into two dimensions using t-SNE and taken on a plane.

5. Discussion and Limitations

The results in Table 1 show that prediction using a weighting function $f_{\text{rank7}}$ is about 10% more accurate than classification using the function $f_{\text{mean}}$, and as the graph shows, we can say that other weight model also outperforms $f_{\text{mean}}$ classification. This indicates that not all problems classified under the appropriate label are similar sentences and that learning all problems may result in noise. In the experiment, we classify the exercises into 24 classes with 115.6 questions per class, and the purpose of the weight function is to determine how many labeled data texts are like the query text. One piece of evidence supporting this is that there are often exercises in the data set that have very similar sentences and are given the same label. Thanks to the presence of such exercises, we were able to achieve a highly accurate multi-class classification even with a small data set.

We also found that the model with n-gram gave about 5% higher accuracy at the classification task, such as $f_{\text{rank7}}$ against SVM. This indicates that a weighted n-gram similarity model is more effective for the classification of short math exercises. In Figure 5 (a)I and (b)I, both models were observed to have regions assigned to units. Also, Figure 5 shows that the word embeddings model’s map has regions assigned to the unit to some extent so that clumps of errors are noticeable in (a)II, but this is not the case for the prediction in (b)II. This indicates that classification using n-gram is more flexible in making predictions.

One limitation of this experiment is that although in this experiment we use two kinds of weighting function $f$ , $f_{\text{mean}}$ and $f_{\text{rank}}$, we can consider other weighting functions. Experimental results are enough to show the better performance of the n-gram representation model than the word embedding model, but we can expect another model yields a more accurate prediction.
6. Conclusion and Future Works

In this study, we propose an automatic classification algorithm that can handle even short sentences in mathematics exercises by focusing on the optimal agreement of a set of mathematical problems. We compared the proposed model which calculates the Jaccard coefficient, to evaluate the similarity of n-gram with the previously existing word embedding model. Experimental results show that the effectiveness in labeling mathematical exercises of the proposed algorithm using weighted n-gram exceeds that of word embedding, with an accuracy of 84.83% versus 79.39%, and an F score of 84.41% versus 79.92% respectively. We also compared the two methods using t-SNE and found that the proposed bi-gram vector space model made more flexible predictions regardless of the position of the vector transformed by t-SNE.

There are some limitations of the current study that should be mentioned, such as how the materials used for the experiment are all from only one publisher which could have lead to the high accuracy in the results. Additional experiments should be carried out in future work to confirm whether the accuracy and F score can be maintained when using materials from various publishers.

Another area that should also be investigated is creating a model that allows multiple labels to be assigned to the same hierarchy and increases versatility for mathematical problems. In the experiment, we preprocessed the data to treat all digit expressions as the same, so in future work experiments should be carried out to check the influence of analyzing mathematical expressions on the classification. We also need to conduct another experiment to quantify the reduction in manual-labeling burden, as well as the difference in the rate of mislabelling from the perspective of the effect on educational consequences.

Moreover, the proposed method needs to address exercises that can be assigned multiple labels. Multi-label classification is also widely used in machine learning (Tsoumakas & Katakis, 2007). There are situations where multiple labels are given because one mathematical problem has multiple knowledge elements.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (Early Career) JP23K17012, (A) JP23H00505, and NEDO JPNP20006.

References


Can We Ensure Accuracy and Explainability for a Math Recommender System?

Yiling DAI\textsuperscript{a}, Brendan FLANAGAN\textsuperscript{b} & Hiroaki OGATA\textsuperscript{a}

\textsuperscript{a} Academic Center for Computing and Media Studies, Kyoto University, Japan
\textsuperscript{b} Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Japan

*dai.yiling.4t@kyoto-u.ac.jp

Abstract: Providing explanations in educational recommender systems are supposed to increase students’ awareness of the recommendations, trust toward the system, motivation to adopt the recommendations. With the expectation to have a higher prediction accuracy, more and more complex recommendation models are developed, which are difficult to explain. It remains debatable that whether there exists a trade-off between the accuracy and explainability of recommender systems. In this study, we focus on the explainable math quiz recommender system—Naïve Concept Explicit (Naïve CE) proposed in our previous work. We are interested in knowing whether the explainable Naïve CE has a good prediction accuracy compared with a powerful but less explainable model—Matrix Factorization (MF). We also proposed a combined model CE+MF to preserve the explainability of Naïve CE and predicting power of MF. We then used a long-term quiz answering dataset to evaluate the models’ accuracy as to predicting students’ correctness rate of the quizzes. The results revealed that 1) The explainable model Naïve CE had a lower accuracy than the less model MF given the sparse dataset; 2) Combining two models achieved a moderate accuracy in predicting students’ answers while preserving the explainability of Naïve CE. Our study served as an example of how to develop an inherently explainable educational recommender system and how to improve the accuracy by integrating more complex models.

Keywords: Recommender system, math quiz, explainability, accuracy, matrix factorization

1. Introduction

To improve learning, educational recommender systems provide recommendations based different criterion including learning activities (Afzaal et al., 2021), knowledge states (Ai et al., 2019; Tang et al., 2019), learning goals (Heras et al., 2020; Huang et al., 2019), learning styles (Heras et al., 2020; Klašnja-Miličević et al., 2011), student profiles (Shanshan et al., 2021) and so on. Unlike consuming entertaining products such as movies or music, reading learning materials or solving quizzes requires higher levels of motivation and cognitive investment of users. Providing explanations of the recommendations to the students is considered as a solution. Some promising results of the explanations’ effects have been found: increasing the students’ attention towards recommended practices, the willingness to open them (Barria-Pineda et al., 2021); increasing the students’ trust in hints, perceived usefulness of them, and the intention to use them again (Conati et al., 2021); increasing the students’ perceived unexpectedness, novelty of recommended courses, and the interests in them (Yu et al., 2021); increasing the students’ usage of the recommended quizzes (Dai, Takami, et al., 2022; Takami et al., 2022).

Unlike traditional machine learning models such as decision tree and logistic regression, recent models involving neural networks are becoming more and more difficult to understand (Khosravi et al., 2022). It is commonly considered that more complex models have better performance in terms of predicting user’s behavior (Molnar et al., 2022; Rudin,
To achieve model transparency, explaining complex models becomes a research field also known as “eXplainable Artificial Intelligence (XAI)” (Arrieta et al., 2020). However, there remain two open issues before suggesting researchers and practitioners to develop more complex models and then struggle to explain it:

1. It is doubted that there is a trade-off between the model’s accuracy and complexity (Molnar et al., 2022; Rudin, 2019). As Gervet et al. (2020) observed in their experiments, the superiority of deep knowledge tracing models over logistic regression models is influenced by the size and shape of the dataset.

2. Should we develop an explainable model in the first place and then improve the accuracy or develop a complex model and try to explain it afterwards (Molnar et al., 2022)?

To address these issues, we focus on a specific educational context of recommending math quizzes in this study. Previously, we proposed a simple and explainable recommender system named Naïve Concept Explicit (Naïve CE) model, which recommends quizzes based on the estimation of the students’ mastery level on math concepts (Dai, Flanagan, et al., 2022). The model also provides explanations of why the students should undertake the quiz from the perspective of math concepts. We also conducted a real-life experiment where the concept-based explanations showed positive effects in motivating students to attempt the quizzes (Dai, Takami, et al., 2022). In this study, we are interested in knowing whether the explainable Naïve CE has a good prediction accuracy compared with a powerful but less explainable model—Matrix Factorization (MF). We also proposed a combined model CE+MF to preserve the explainability of Naïve CE and predicting power of MF. We used a long-term quiz answering dataset in our learning management system to evaluate the models’ accuracy as to predicting students’ correctness rate of the quizzes. The results revealed that 1) The explainable model Naïve CE had a lower accuracy than the less model MF given the sparse dataset; 2) Combining two models achieved a moderate accuracy in predicting students’ answers while preserving the explainability of Naïve CE. Our study served as an example of how to develop an inherently explainable educational recommender system and how to improve the accuracy by integrating more complex models.

2. Related Work

2.1 Explainability of Recommender Systems

Basically, there are two approaches to generate explanations in recommender systems—model-intrinsic and post-hoc (Zhang & Chen, 2020). In the model-intrinsic approach, the explanation explains exactly how the model generates a recommendation. In educational contexts, an example of model-intrinsic explanation can be to explain how the student’s knowledge state is estimated and why a learning item is considered preferable to improve his/her knowledge state (Dai, Flanagan, et al., 2022). Other model-intrinsic explanations include rule-based (Conati et al., 2021), keyword-based (Yu et al., 2021), concept-based (Dai, Flanagan, et al., 2022; Rahdari et al., 2020), and parameter-based (Takami et al., 2022) explanations. In contrast, the post-hoc approach allows the recommending mechanism to be a “black box” and generates the explanations afterwards. In educational contexts, a post-hoc explanation for a recommended item can be something not necessarily related to the knowledge state estimation but instrumental in motivating the student to accept the recommendation. For instance, an explanation showing how many students have attempted this item may work for students who are weak to peer pressure (Takami et al., 2023). Feature-based explanations were adopted for “black-box” models but Swamy et al. (2022) found that the explainers are not consistent on feature importance. “Black box” or “Deep” methods are commonly considered more powerful in predicting users’ behavior (Molnar et al., 2022; Rudin, 2019). However, this remains doubtful as Gervet et al. (2020) found that deep knowledge tracing models worked better with larger datasets while logistic regression models worked better with denser datasets. There is a concern of over-using
complex models while an explainable model which has a comparable accuracy is available (Khosravi et al., 2022; Molnar et al., 2022).

In the context of learning math, we consider that developing an explainable recommender system in the first place is intuitive and straightforward. We proposed a simple and explainable recommender system named Naïve Concept Explicit (Naïve CE) model, which recommends quizzes based on the estimation of the students' mastery level on math concepts (Dai, Flanagan, et al., 2022). We also conducted a real-life experiment where the concept-based explanations showed positive effects in motivating students to attempt the quizzes (Dai, Takami, et al., 2022). In this study, we aim to investigate the predicting accuracy of Naïve CE compared with other classic recommendation models. We also attempt to improve the accuracy of Naïve CE by combining it with more complex models while preserving the explainability. This study serves as an example of developing inherently explainable recommender system as suggested by Molnar et al. (2022).

### 2.2 Naïve CE and MF for Recommending Math Quizzes

As Birenbaum et al. (1993) suggested, identifying specific misconceptions and difficult areas is more instructive than a test score for remediation in learning Algebra. Therefore, it is important to recommend math quizzes that address the students' weak points which are readable math concepts. This motivated us to propose a concept-explicit recommender system named Naïve CE (Dai, Flanagan, et al., 2022). Naïve CE assumes that solving a quiz requires the knowledge of related math concepts in the quiz. We utilized the student-quiz interactions and quiz-concept associations to estimate students' mastery levels of the concepts and the possibilities for them to answer the quizzes correctly. We then recommended the quizzes based on the possibilities and the expected learning gains in terms of the mastery level updates of the concepts. Naïve CE is inherently explainable as every step of the estimation is a shallow calculation which is understandable for human beings.

However, we also have a concern on the estimation performance of Naïve CE. In other words, how well can Naïve CE estimate students' mastery level and their probabilities to correctly answer the quizzes? Is there an inferiority in estimation performance compared with more accurate but less explainable models? As Barnes (2005) pointed out, it is debatable whether an explicit model with expert-assigned concepts models student performance better than an implicit model with latent factors. Therefore, we selected a classic recommendation model matrix factorization with latent factors as a comparative target of Naïve CE.

Matrix factorization (MF) (Koren et al., 2009; Takács et al., 2008) is a frequently used model to recommend items that users may have interests. This model assumes a user's interest towards an item comes from her/his preferences on some factors and the relatedness of the factors with the item. It then guesses the unseen user-item interactions by learning from observed user-item interactions. Khosravi et al. (2017) applied MF in their model to estimate students' knowledge gaps to answering the quizzes. To integrate the strengths of both models, Abdi et al. (2018) fed the error in Bayesian knowledge tracing model to an MF model, improving the accuracy of estimating student performance. Since MF has been verified as a useful model to estimate student performance and shares some similarities in modeling the problem with Naïve CE, we chose MF as a comparative model in discussing the estimation performance and model explainability. We also propose a method to combine two models so that the readability of concepts in Naïve CE is preserved.

### 3. Math Recommender Models

#### 3.1 Problem Definition

In learning management systems, the learning activity can be modeled as a sequence of students' reactions towards learning materials. The task is to recommend learning materials that fit to an individual student's learning progress. It is common that the observed student...
reactions are limited to a small set of the learning materials. As a result, predicting the student reactions on unseen learning materials is a key step. For the specific context of solving math quizzes, we formalize the problem as follows: Given a set of \( m \) students, a set of \( n \) math quizzes, and the student correctness rates on the quizzes \( R \in \mathbb{R}^{m \times n} \), we want to estimate the student correctness rates on unseen quizzes \( \hat{R} \in \mathbb{R}^{m \times n} \).

3.2 Naive Concept-Explicit Model (Naïve CE)

Suppose that we have a quiz “Let the set of all positive divisors of 12 be A. Fill in the \( \square \) with \( \varepsilon \) or \( \varepsilon \). (1) 2\( \square \)A (2) 7\( \square \)A (3) 12\( \square \)A”. Solving the math quiz requires the knowledge of “set” and “positive divisor”. The probability that a student can successfully solve a math quiz depends on how s/he understands the required concepts. Motivated by this intuition, we proposed Naïve CE (Dai, Flanagan, et al., 2022) and we review the mechanism of the model as shown in Figure 1:

**STEP 1** When given the observed student-quiz matrix whose entries indicate the correctness rates and the quiz-concept matrix whose entries indicate the relatedness of a concept and a quiz, we calculate the students’ mastery level on each concept by looking at how they successfully solved quizzes related to the concept. Note that the quiz-concept correspondence is extracted from the quiz information automatically, which is also readable concepts to students.

**STEP 2** We then estimate the probability of a student successfully solving a quiz by considering how much of the required concepts has been mastered.

By doing this, the probabilities are modified by the inter-relationships between quizzes and concepts. For instance, \( s_1 \) successfully solved \( q_1 \) in the history but got an estimated success of 0.67. This is because \( q_1 \) requires the knowledge of \( c_1 \) and the student failed to solve \( q_2 \) which also requires the knowledge of \( c_1 \). However, this model falls short in coping with unseen concepts. For instance, \( s_2 \) had not attempted any quizzes related to \( c_3 \). As a result, \( c_3 \) is ignored in STEP 2.

3.3 Matrix Factorization (MF)

MF decomposes the observed user-item interaction matrix \( R \in \mathbb{R}^{m \times n} \) into two matrices \( P \in \mathbb{R}^{m \times k} \) and \( Q \in \mathbb{R}^{k \times n} \) such that \( R \approx PQ \), where \( k \) is the number of latent factors. By minimizing the difference between the estimated and the observed interactions (also viewed as a machine learning process), we get full \( P \) and \( Q \), which help us estimate the unseen interactions. The magic part of this model is that it supposes a user's interest on an item comes from the synergistic effect of the user's preferences on the latent factors and the importance of the factors to the item. This idea is somehow similar to Naïve CE except that the factors are “latent” and difficult to interpret. There is a variant of MF which considers user bias and item bias. As a user may tend to highly rate all items or an item of low quality tends to be rated low by all users, introducing bias parameters in MF helps model this situation. By doing this, the sum of \( PQ \) and bias better approximates the observed interactions but the intermediate value of \( PQ \) is harder to interpret.
3.4 Concept-Explicit Matrix Factorization (CE+MF)

In Section 3.2, we discussed that Naïve CE is easy to interpret since the concepts are predefined and the computation is simple and straightforward. However, it simply gives up guessing when encountering unseen concepts and quizzes. In Section 3.3, we described that MF is good at estimating the probable values for unseen interactions by iteratively learning from the observed interactions. However, the latent factors of the consisting matrices are difficult to interpret. To take the strengths of both models, we propose a simple hybrid model called CE+MF. As illustrated in Figure 2, Naïve CE utilized the observed student-quiz matrix and quiz-concept matrix to estimate student-concept matrix. The student-concept matrix is again used to adjust the student-quiz matrix. MF simply decomposes the observed student-quiz matrix into two matrices with latent factors. In CE+MF, we first estimate the student-concept matrix as we do in STEP 1 in Naïve CE model. Then, we adopt MF model to update the student-concept matrix where the mastery level on unseen concepts is modified. Last, we update the student-quiz matrix with the updated student-concept matrix as we do in STEP 2 in Naïve CE model. As a result, CE+MF model is supposed to have a higher predictive performance than Naïve CE model while preserving the explainability on concepts.

4. Evaluation

In this study, we aim at investigating two aspects of the recommender models—-quiz mastery estimation performance and explainability. For the quiz mastery level estimation performance, we use a historical dataset collected in a learning system to evaluate whether the models can correctly predict students’ answers for unseen quizzes. For explainability, we evaluate whether each step in the model can be explained and what user-friendly explanations can be provided.

4.1 Quiz Mastery Level Estimation Performance

4.1.1 Dataset

We collected quiz answering data from our learning system (Flanagan et al., 2021) generated by the first-year students of a Japanese high school from April 2021 to March 2022. During this period, the students attempted the math quizzes in different contexts such as finishing the assignments, preparing an upcoming test, and self-oriented practicing. As they attempted a quiz, they were required to check the answer and report whether they solved the quiz successfully. Each attempt was recorded as a 0-1 score associated with the student id, quiz id, and timestamp. We computed the aggregated student-quiz correctness rate by taking the average score of all attempts throughout the period. We did not conduct any data filtering process as the temporal order of attempts and the number of attempts are not essential in this evaluation framework. Finally, we obtained a dataset consisting of 27,431 attempts for 270 unique students and 1,919 unique quizzes. Table 1 shows the

![Figure 2. The mechanism of CE+MF.](image)
statistics of the number of attempts per student and per quiz. After converting the log data into the student-quiz correctness matrix, only 23,155 pairs of students and quizzes were observed, which indicates a very high sparsity of 95.53% ($1 - \frac{\text{# observed pairs}}{\text{# students} \times \text{# quizzes}}$).

Table 1. Statistics of the Dataset for Quiz Mastery Level Estimation Evaluation

<table>
<thead>
<tr>
<th></th>
<th># attempts per student</th>
<th># attempts per quiz</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>101.596</td>
<td>14.294</td>
</tr>
<tr>
<td>std</td>
<td>97.681</td>
<td>31.404</td>
</tr>
<tr>
<td>min</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>max</td>
<td>808</td>
<td>426</td>
</tr>
</tbody>
</table>

4.1.2 Metrics

Our main concern is to evaluate whether a model can predict a student’s success probability on a quiz. Therefore, we adopt two metrics to measure the agreement between the estimated probability and true correctness rates: **Area under ROC curve (AUC)** is considered an effective metric to measure how well a model separate negative and positive samples across different decision threshold choices (Bradley, 1997). Since the true correctness rates for student-quiz pairs are real numbers between 0 to 1, we first transform the true correctness rates into 1 if it is greater than 0.5, 0 otherwise when applying AUC. **Root mean square error (RMSE)** is used to measure the absolute differences between the estimated probability and the true correctness rates.

4.1.3 Implementation

As illustrated in the upper part of Figure 3, we set aside 20% of the student-quiz attempts as test data and all models were blind to these data during training or tuning process. For models involving hyper-parameter tuning, we adopted a 5-fold cross validation approach to select the best combination of parameters. As illustrated in the lower part of Figure 3, in each fold, 20% of the data is used to validate the model performance and the average performance of all the folds is treated as the final performance of a combination of parameters. We adopted a grid-search approach to generate the combinations of parameters. Only the performance of the best combination of parameters will be reported in the following section.

The followings are some implementation details of the models: 1) Naïve CE. As described in (Dai, Flanagan, et al., 2022), we adopted text mining techniques to automatically extract math concepts from the quizzes. The entries of quiz-concept matrix were computed using TFIDF (Salton & Buckley, 1988) weighting scheme. 2) MF. We adopted stochastic gradient descent algorithm to obtain the matrices $P$ and $Q$ whose product has the minimum difference with the observed student-quiz correctness rate. We tuned three parameters--- learning rate $\alpha$, regularization factor $\gamma$ and number of latent factors $k$ for MF. MF_bias and CE+MF_bias are variants with bias parameters of MF and CE+MF.
respectively. The best combination of hyperparameters is $\alpha = 0.01, \gamma = 0.1, k = 5$ and is used in all MF-related models. The code for MF-related parts was adapted from (Yeung, 2020).

### 4.1.4 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve CE</td>
<td>0.639</td>
<td>0.508</td>
</tr>
<tr>
<td>CE+MF</td>
<td>0.688</td>
<td>0.478</td>
</tr>
<tr>
<td>CE+MF_bias</td>
<td>0.692</td>
<td>0.464</td>
</tr>
<tr>
<td>MF</td>
<td>0.772</td>
<td>0.419</td>
</tr>
<tr>
<td>MF_bias</td>
<td>0.799</td>
<td>0.381</td>
</tr>
</tbody>
</table>

Table 2 shows the AUC and RMSE values for each model. Overall, Naïve CE has the lowest, MF has the highest, and CE+MF has the medium performance in both metrics. MF with bias has a better performance both in separate and hybrid models. This result is consistent with our expectation:

- Naïve CE is straightforward but ignores the unseen concepts or quizzes. From the perspective of AUC, this model can discriminate a correct or incorrect answer. However, the performance is nearly a random model from the perspective of RMSE, which means the detailed values of the correctness has a large gap with student's true mastery level.
- MF and MF_bias do a good job at approximating the observed student-quiz correctness rates and therefore the latent factors help to predict the values for unseen pairs. With the AUC value being close to 0.8, MF_bias is supposed to be practically useful to separate a correct or incorrect answer (Mandrekar, 2010). Meanwhile, the RMSE value is still high if we consider a situation where we mistake a student's correctness rate 0.98 into 0.6. However, whether the students can recognize the difference and how they perceive the estimation needs to be further investigated and discussed.
- The hybrid model CE+MF achieves better performance than Naïve CE but still has a distance to the one of MF. We consider a possible reason is the flaws in quiz-concept matrix. First, not all necessary knowledge and skills for solving a math quiz can be detected from the textual information of the quiz. Second, the relatedness of the concepts to a quiz may not be correct just judging from their occurrences in the quiz. Ingesting more elaborated domain models to Naïve CE and observe the performance improvement is one of the future directions.

### 4.2 Explainability

We compare the explainability of different models as shown in Table 3:

- At the lowest level, all the models can provide information about the estimated student-quiz correctness rate, which can be used to indicate the difficulty of the quiz when recommending the quiz.
- At the medium level, we try to further explain why the correctness rate is as it is. Naïve CE, CE+MF, and CE+MF_bias can provide information about the students' mastery level of concepts, which is the rationale behind the student-quiz correctness rate estimation. However, MF suffers from provide human-readable information about how the student-quiz correctness rate was estimated as the factors are latent. Fortunately, we can leverage the user bias and item bias in MF_bias to provide some extra information. Specifically, we treat the user bias as the student's general ability to solve math quizzes and the item bias as the quiz's general difficulty to all students. Note that it is potential to improve the accuracy of Naïve CE by introducing quiz bias or student bias, which will be future work.
- At the highest level, we want to explain why the concept master level or quiz general difficulty is estimated. In Naïve CE, the concept mastery level is explainable as the quiz-
concept associations is readable to human. In contrast, CE+MF or CE+MF\_bias involves a matrix factorization process in estimating the concept mastery level, which is difficult to explain.

Table 3. Explainability and Explanations of Recommender Models.

<table>
<thead>
<tr>
<th>Explainability</th>
<th>Model</th>
<th>Argument</th>
<th>User-Friendly Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Naïve CE</td>
<td>student-quiz correctness rate</td>
<td>The estimated difficulty of this quiz for you is 70%.</td>
</tr>
<tr>
<td></td>
<td>MF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MF_bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE+MF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE+MF_bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Naïve CE</td>
<td>concept mastery level</td>
<td>Your mastery level of these concepts can be improved by solving this quiz: multiple, integer, proof</td>
</tr>
<tr>
<td></td>
<td>CE+MF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE+MF_bias</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MF_bias</td>
<td>quiz general difficulty</td>
<td>This quiz is difficult (84%) for most of your classmates. Let's have a try!</td>
</tr>
<tr>
<td></td>
<td>MF_bias</td>
<td>student general ability</td>
<td>Your ability (23.4%) to solve quizzes is lower than the average (56.9%) of your classmates.</td>
</tr>
<tr>
<td>High</td>
<td>Naïve CE</td>
<td>quiz-concept associations</td>
<td>Integer is important in solving q1 (50%), q2 (30%), and q3 (20%). Since you have mistaken q1 twice, you are suggested to address this problem first.</td>
</tr>
</tbody>
</table>

To summarize, Naïve CE is explainable until the highest level as every step in the model is a shallow computation from observed data. MF is only explainable at the lowest level as the latent factors are difficult to interpret. However, the variant MF\_bias possesses some additional information about the general information of the quiz difficulty and student ability. This shows a direction to improve the accuracy of Naïve CE by introducing quiz difficulty and student ability parameters. CE+MF is explainable at the medium level as the whole framework is identical to Naïve CE, but a local step involves a MF process.

5. Discussion

As the results in Section 4.1 show, MF has the highest performance of estimating quiz mastery level while Naïve CE has the lowest, CE+MF has the medium performance. The results in Section 4.2 show that we can preserve part of the explainability of an inherently explainable model Naïve CE by combining it with a difficult-to-explain model MF. Given the fact that the quiz answering data set is sparse and the quizzes only have 14 answers on average, we did observe a trade-off between the accuracy and the explainability of different recommender models. As was explored in Gervet et al.’s work (2020), models’ performance varies to the type and characteristics of the dataset and features. We think it is important to select a proper recommendation model based on the learning context and data available. Besides, we want to clarify that the explainability of the model is not necessarily equal to the explainability to the students in the practical world. For instance, it would be sufficiently explainable if the students are satisfied with the quiz general difficulty without the interest to understand how it is computed. In this case, MF\_bias is not easy to explain in the sense that every step is understandable to humans but explainable to end users.

6. Conclusions and Future Work

In this study, we focused on exploring the accuracy and the explainability of math recommender systems. We took a simple and explainable recommender model Naïve CE as an example and compared its performance with a more complex but difficult-to-explain
model MF. We also attempted to combine two models so that the strengths of both models are integrated. Using a student quiz answering dataset, we found that the explainable model Naïve CE had a lower accuracy than the complex model MF, and the combined model CE+MF had a moderate accuracy. Our results also showed that it is possible to improve the accuracy of an inherently explainable model and preserve the explainability by combining it with more complex models.

Some directions of future work are: 1) Explore other models that can be integrated into the framework of Naïve CE and introduce other learning related parameters such as quiz difficulty and student ability; 2) Investigate how dataset influences the performance of the recommender models; 3) Ingesting more elaborated domain models into Naïve CE and explore the performance improvement; 4) Further explore the difference between the model explainability and the practical explainability to students. Students with different levels of motivations, information literacy, and curiosity may question the recommendations at different levels. It may be important to personalize the explanations according to different scenarios.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

References


Activity Analysis Support System by Causal Relation Check Between Sub-activities

Kota KUNORI\textsuperscript{a} \& Tomoko KOJIRI\textsuperscript{b}
\textsuperscript{a}Graduate School of Science and Engineering, Kansai University, Japan  
\textsuperscript{b}Faculty of Engineering Science, Kansai University, Japan  
*k194388@kansai-u.ac.jp

Abstract: When we encounter problems while performing specific activities, we need to identify the cause of the problem to fix it. An activity is composed of several small activities (sub-activities) and the cause might lie in either one of them. We therefore need to analyze the components of the activity to identify the cause. However, since valid sub-activities are not defined, it is difficult to perform a sufficient analysis of the activity. The objective of this research is to propose a method for analyzing an activity and identifying the problematic sub-activity. To attain this goal, valid sub-activities should be defined. This research proposes an activity model that represents the relations between an activity and its sub-activities. In addition, it constructs an activity analysis system with functions to provide hints for deriving sub-activities and verify the validity of the derived sub-activities.

Keywords: Problem analysis, activity model, validity check

1. Introduction

Project Based Learning (PBL) has often been adopted for cultivating problem-solving skills. In PBL, students acquire knowledge and skills by finding and solving problems in small groups. However, sometimes these activities are not practiced successfully because the student lacks the skill to practice the activities. There are several studies that support PBL; many focus primarily on constructing the groupware that facilitates communication by enabling the exchange of products and opinions within the group (Rizal, Rusdiana, Setiawan, \& Siahaan, 2021) (Chen, et al., 2021). However, most do not support the learning of the problem-solving skill itself. Problem solving is accomplished by analyzing the problem, formulating a solution, and implementing it, so support at each step is necessary. This study focuses on analyzing the problem and develops a support system for problem analysis.

Some studies support the identification of the cause of a problem by presenting a method for analyzing the problem. Harich et al. proposed a worksheet that enables group problem solving by repeatedly asking the question "Why?" until a consensus is reached within the group (Harich \& Rosas, 2020). The problem analysis methods presented in these studies can derive the cause of a problem, but do not guarantee the validity of the derived cause. For the derivation of valid causes, there are studies in which the system takes the place of the derivation or adaptively supports the derivation based on problem domain knowledge. For instance, Mori et al. focused on analyzing the research problem of the educational technology domain (Mori, Hayashi, \& Seta, 2019). They constructed an ontology for conducting such research and proposed a system for generating questions that allow the user to consider the elements based on the ontology. Since these studies support valid analysis based on the system's problem domain knowledge, these studies can only support problem analysis within the scope of the system's problem domain knowledge. Since there are many different types of everyday problems, it is not possible to describe all the problem domain knowledge in advance. We must propose a generic method of problem analysis rather than have the system carry the knowledge of the problem domain.

Problem analysis is the activity of understanding the structure of the problem and discovering its causes. In the analysis, the elements that compose the problem should be derived. However, it is difficult to derive enough elements that can form the problem.
Therefore, we define the conditions of what should be derived in the problem analysis. We also clarify what needs to be considered to derive elements that satisfy the conditions and construct an analysis support system that leads the analysis and judges whether the analyzed elements satisfy the conditions.

Among the many types of problems we face in daily life, this research focuses on problems that occur during activities—problems that interfere with an activity’s successful performance. Therefore, the problem analysis is to derive the sub-activities of the activity.

2. Activity Model

An activity is an action sequence taken to achieve a goal and is composed of several sub-activities. Figure 1 shows the activity model. Activities have a condition that indicates the state in which they can be performed. The difference between the pre-state and post-state can be regarded as the sub-activity’s effect.

The sub-activity sequence can achieve the same activity as the original activity, so the sub-activities must have the same effect as the original activity. In addition, the sub-activities should have a causal relation with each other where the post-state of the causal activity becomes the condition for the resulting activity.

3. Activity Analysis Support System

This study constructs a system to support user analysis of activities based on the activity model. Figure 2 shows the configuration of the system.

It consists of an interface, a validity judgment function, and a hint function.

Imagining the scenes in which the activities are actually performed is effective for deriving the sub-activities. Therefore, our interface lets the user describe the activity/sub-activities with the scene.

The validity judgment function checks whether the derived sub-activities satisfy the conditions. It consists of an effect sufficiency judgment function that judges whether all candidate sub-activities have an effect, a coherence judgment function that checks whether all sub-activities are connected by a causal relation, and a consistency judgment function that determines whether the effect of the entire sub-activities sequence is consistent with that of the original activity. The hint function suggests what to consider to derive a valid sub-activities sequence.

The user can input the activity to be analyzed, the candidate sub-activities, and the effects and conditions of the activity/candidate sub-activities into the interface. When the user requests a validity judgment of the sub-activities, the validity judgment function judges whether the candidate sub-activities entered by the user are valid or not and provides reasons if it is not valid. When the user requests a hint recommendation, the hint function recommends the appropriate operation based on the validity judgment result.
Figure 3 shows the system interface, consisting of an activity display area and an operation area. In the operation area, the user can input the activity, sub-activity, and effect and condition that the user derived in his or her analysis. The activity display area shows the inputted sub-activities and their causal relations. The sub-activity is shown in table format with the name of the sub-activity on top and its effect and condition below. The causal relation is represented by the arc between sub-activities. The system searches two sub-activities, one whose condition is the same as the effect of the other, and, if found, draws an arc between them.

The analysis support panel in the operation area contains two buttons: a button that asks to check the validity of the input sub-activities and a button that requests a hint. When the validation check button is pressed, it displays a message indicating that the sub-activity is valid if it is judged to be valid. If not, it displays the reason why it is not valid. When the hint button is pressed, it displays a message telling the user what to do next.

4. Conclusion

This paper proposes an activity model and the activity analysis support system based on it. The activity analysis support system provides an interface that lets the user consider not only the sub-activity but also its condition and effect. It also provides a function that judges whether the candidate sub-activity sequence is valid and gives hints that inform the user what to consider to derive a valid sub-activity sequence. As our next step, we will conduct experiments to evaluate the validity of the activity model and the effectiveness of the proposed system.

References


Method for Estimating Learning Strategies from Tools Using Bayesian Network

Kento KUWAJIMA*, Atsushi ASHIDA & Tomoko KOJIRI

a Graduate School of Science and Engineering, Kansai University, Japan
b Institute for Datability Science, Osaka University, Japan
c Faculty of Engineering Science, Kansai University, Japan
* k268505@kansai-u.ac.jp

Abstract: With the recent development of the robot technology, teacher robots will be introduced into educational settings to instruct learners. Such teacher robots, like human teachers, will be required not only to teach knowledge to learners, but also to monitor their learning process and motivate them. Positive learning attitudes can be ascertained by changing the learning strategy. The fact that learners change their learning strategies may indicate that they are making efforts to find an appropriate strategy. The learning strategies used by learners can be inferred from the tools they use to learn. Therefore, we propose a method to estimate learning strategies using Bayesian networks when learners change them.

Keywords: learning strategy, teacher robot, motivation, Bayesian Network

1. Introduction

In the future, it is expected that robots will play an increasingly important role in education, such as to provide lessons to learners. Such teacher robots will be required to provide instruction equivalent to that of a human teacher.

When human teachers teach, they not only impart knowledge to learners, but also they often motivate learners to study. One of these efforts is to provide instruction that increases the learner’s motivation. In the future, such instruction will also be required for teacher robots. However, current robot-based instruction focuses only on knowledge teaching (Ono et al. 2019, Tanaka & Matsuzoe, 2012), and there are few examples of instruction that is oriented toward motivation. Therefore, this paper proposes a method for robot teachers to enhance learners' motivation.

One of the methods for enhancing the motivation is to evaluate the learner’s effort and ingenuity in the learning activity and praise them. The effort and ingenuity are often appeared as changes in learning methods during the learning process, so it is required to grasp the changes in the learner's learning methods.

For a robot teacher to read changes in a learner's learning method, it is necessary to estimate changes in the learning method from what the robot can observe. Therefore, the research questions for this study are as follows:

- What observable information can be used to estimate changes in learning methods?
- How to estimate learning methods using these observables tools?

Usually, learners use learning tools, such as notes, pencils, and rulers, and such tools are fixed according to the learning methods. Based on this consideration, this study focuses on the learning tools as observable information of the robot and proposes a method for estimating changes in the learning method from them.

2. Overview of Learning Method Estimation System

The learning method is a pair of learning strategies and learning means. A learning strategy is a way for learners to approach learning in order to achieve their learning goals. An example of strategies is "memorizing" when the learning goal is to learn English words. In order to
practice the learning strategies, the learner selects a learning means, such as "writing repeatedly" for the strategy "memorize." Many of the learning means use tools. For example, to achieve the means of "writing repeatedly," a pencil that can be used for writing, and to achieve the means of "divide into sections," a multi-color pen that can mark the segmented items in different colors might be used as tools. Therefore, the learning method, which is a pair of strategies and means, is characterized by the tools.

In this study, for the teacher robot to estimate the learning method, Tools-means-strategies model, which represents the relationship between tools, means, and strategies, is introduced. By observing the tools through the sensor, the learning means and the learning strategies that are currently taken are estimated by using Tools-means-strategies model.

As the first step of realizing such teacher robot, we construct a system that corresponds to a function that identifies the current learning means and learning strategy. An overview of the proposed system is shown in Figure 1. The system has Tools-means-strategies model and a learning record that stores the history of the learner's learning strategies. When the user of the system inputs the tools that a learner used for the learning, instead of observing it using the sensors, the learning method identification function estimates the learning strategy based on Tools-means-strategies model and stores it as a learning record.


3.1 Tools-means-strategies Model

Tools-means-strategies model proposed applied Bayesian networks to represents the causal relationship between the strategy and the means, and between the means and the tool. A Bayesian network represents causal inference by means of a directed acyclic graph structure, where nodes represent variables and links represent probabilistic correlations among variables. Each node has a corresponding choice, each of which has a probability that the choice will be chosen when a superior node makes a decision. Each alternative within the same node is exclusive and does not occur simultaneously.

In the Tools-means-strategies model, the nodes are the strategy, the means, and the tool, and the links exist between the strategy and the means and between the means and the tool. When multiple tools are used at the same time, the set of tools are regarded as a single tool. Figure 2 shows an example of Tools-means-strategies model for English vocabulary learning.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Memorize</th>
<th>Analyze the structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divide into sections</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Write repeatedly</td>
<td>0.9</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tools/Means</th>
<th>Tools-means-strategies model</th>
<th>Learning method identification function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pencil Notebook</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Multi-color pen Dictionary Notebook</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Figure 2. Example of Tools-means-strategies Model
3.2 Learning Method Identification Function

The learning method identification function uses the Tools-means-strategies model to estimate the current strategy of the learner based on the input tools. The probability of each strategy is calculated from the model, and the strategy with the highest calculated value is determined to be in use. The probability that a learner is using a certain strategy A when a tool B is observed is calculated by the following formula.

\[
(Probability \ that \ a \ learner \ uses \ strategy \ A \ when \ a \ tool \ B \ is \ observed) = (Probability \ of \ using \ strategy \ A) \times \sum_{all \ means} \{(Probability \ of \ using \ each \ means \ when \ using \ strategy \ A) \times (Probability \ of \ using \ tool \ B \ when \ using \ each \ means)\}
\]

Let’s consider the example using the Tools-means-strategies model shown in Figure 2. If a pencil and a notebook are observed, the probability using each strategy can be obtained by normalizing the values obtained in the following equation so that the sum is 1.

Memorize: \(0.8 \times (0.1 \times 0.7 + 0.9 \times 0.9) = 0.704\)
Analyze the structure: \(0.2 \times (0.4 \times 0.7 + 0.6 \times 0.9) = 0.164\)

Normalizing the obtained values and rounding to the fourth decimal place, the probability of using the strategy of memorizing when using a pencil and notebook is 0.811 and the probability of using the strategy of analyzing the structure is 0.189, which indicates that the learner is taking the strategy of memorizing.

4. Conclusion

In this study, we proposed a method for estimating a learner’s learning strategy from the tools used by the learner. Since there is a causal relationship from strategies to means and from means to tools, Tools-means-strategies model was introduced that can infer strategies based on the causal relationships when tools are observed. Currently, we have constructed the model for English vocabulary learning based on the results of a questionnaire survey asking the strategies, means, and tools in the English vocabulary learning. In future, we need to evaluate the validity of this constructed model. In addition, we also need to develop other functions of the teacher robot.

The current system treats tools as concrete instruments. However, just as notebooks are being replaced by PC-based document creation tools in line with the recent digitalization, the tools used may change in the future. Tools have functions, and tools with the same functions are likely to be used in the same way and in the same manner. For example, a pencil and a tablet pen both have the function of writing, and a ruler and a triangular ruler both measure straight lines. By expressing the tools by their function, not by the name, Tools-means-strategies model can be used even if tools change in the future. Therefore, we need to modify the tool nodes in Tools-means-strategies model to represent the their functions, not to their names.

References

ExGen: Ready-To-Use Exercise Generation in Introductory Programming Courses

Nguyen Binh Duong TA*¹, Hua Gia Phuc NGUYENa & Swapna GOTTIPATIa
¹School of Computing and Information Systems, Singapore Management University
*donata@smu.edu.sg

Abstract: In introductory programming courses, students as novice programmers would benefit from doing frequent practices set at a difficulty level and concept suitable for their skills and knowledge. However, setting many good programming exercises for individual learners is very time-consuming for instructors. In this work, we propose an automated exercise generation system, named ExGen, which leverages recent advances in pre-trained large language models (LLMs) to automatically create customized and ready-to-use programming exercises for individual students on-demand. The system integrates seamlessly with Visual Studio Code, a popular development environment for computing students and software engineers. ExGen effectively does the following: 1) maintaining a set of seed exercises in a personalized database stored locally for each student; 2) constructing appropriate prompts from the seed exercises to be sent to a cloud-based LLM deployment for generating candidate exercises; and 3) implementing a novel combination of filtering checks to automatically select only ready-to-use exercises for a student to work on. Extensive evaluation using more than 600 Python exercises demonstrates the effectiveness of ExGen in generating customized, ready-to-use programming exercises for new computing students.

Keywords: introductory programming courses, exercise generation, large language models, prompt engineering, auto-filtering.

1. Introduction

Despite recent advances in code generation using large language models (LLMs), e.g., OpenAI Codex, ChatGPT, Google LaMDA, etc., programming is still an essential skill that computing students must master in the foreseeable future (Becker et al., 2023). It is well-known that computer programming is a challenging subject for many new university students (Keuning et al., 2018). To learn effectively, students must practice on their own frequently with suitable exercises designed for their level of programming skills and knowledge. However, it is usually impractical for instructors to manually create enough exercises for all students, let alone creating customized or personalized exercises for each individual student on-demand, i.e., when they need to practice. Exercises from Internet platforms for coding practices and interview preparations (Joshi et al., 2023) such as LeetCode.com might not be ready-to-use in introductory courses as they are not customized for the skill levels of new computing students. From our experience, many easy-level LeetCode problems can be quite challenging for novice programmers (Ta et al., 2022). Intermediate or harder problems found online often require advanced data structures and algorithm concepts not taught in an introductory course.

Automatic generation of programming exercises has been an important problem in technology-enabled education (Zavala and Mendoza, 2018). Recently, the application of pre-trained LLMs in supporting exercise and feedback generation has been gaining attention. Sarsa et al. (2022) explored OpenAI Codex for the purpose of creating new Python exercises and code explanations. They provided a comprehensive, mostly manual evaluation of the output generated by Codex. They showed that a large percentage of the generated exercises were sensible and new, but most of them were not directly usable by students. However, they did not address the problem of on-demand generation of ready-to-use exercises customized
for various levels of difficulty. Recently, Kazemitabar et al. (2023), Becker et al. (2023) and Finnie-Ansley et al. (2023) discussed and evaluated the use of LLMs to generate sample solutions and to provide adaptive feedback on code submitted by students, but not exercise generation.

In this work, we propose ExGen, a software tool for automatically generating new, ready-to-use exercises for individual students learning introductory programming in Python. ExGen makes use of LLMs which have been pre-trained on a massive amount of text and code such as OpenAI's GPT-3.5-Turbo. In ExGen, we implement carefully designed LLM prompting strategies, automatic filtering of exercises, and seamless integration with Visual Studio (VS) Code. ExGen makes it convenient for students to work on new exercises that are customized for their skill levels. An innovative component of ExGen is that it makes use of LLMs not just to generate new exercises, but also to automatically check if the newly generated exercises are ready-to-use or not. In this way, ExGen is different from existing research that investigated exercise generation based on pre-defined templates (Zavala and Mendoza, 2018). We make the following contributions in this paper:

- We considered the problem of auto-generating ready-to-use exercises customized for a certain difficulty level and programming concepts, e.g., string, lists, etc., upon requests from individual students.
- We designed zero-shot and few-shot prompting strategies to obtain appropriate responses from pre-trained LLMs. We then proposed a mechanism of chaining different filtering checks which auto-select ready-to-use exercises instead of requiring human experts to inspect the LLM output and make decisions. The checks are based on popular software engineering techniques such as unit testing combined with LLM-based checking techniques.
- We implemented a fully functional VS Code extension which makes it more convenient for students to generate new exercises and practice coding. We noted that using existing web interfaces such as ChatGPT or OpenAI's Playground requires a lot of copy/paste actions for adjusting prompts and submitting requests.
- We conducted an extensive evaluation of ExGen using more than 600 programming exercises generated via the GPT-3.5-Turbo API. Each of these exercises was manually inspected to see if they are ready-to-use; and compared to ExGen's output. The result demonstrated the effectiveness of ExGen and the proposed solution design.

2. Problem Statement

We are interested in auto-generating ready-to-use exercises, i.e., those that a student can work on right away without further modifications by human experts. To be considered ready-to-use, a generated exercise would need to have: 1) a clear problem statement, 2) working solutions and test cases, and 3) an appropriate difficulty level as requested by the student. This is a challenging problem, as Sarsa et al. (2022) found that most generated exercises were not ready-to-use when using OpenAI Codex which is based on the GPT-3 model.

On top of generating clear, well-defined problem statements, producing exercises with the right level of difficulty is of crucial importance to students learning programming. For instance, beginners would be encouraged by easy exercises covering the concepts taught in class as they are just getting started. On the other hand, above average students would like to be challenged with harder problems which require a certain level of higher order thinking. Many programming problems available on Internet platforms (Joshi et al., 2023) are not appropriate for students in our course due to them: 1) being very difficult, and/or 2) requiring advanced concepts such as dynamic programming, tree-based data structures, etc., which are not covered in an introductory course for first-year students.

In this work, we define three levels of difficulty for exercises which have been used in our introductory programming course at Singapore Management University. The three levels of difficulty, namely "easy", "intermediate", and "hard", are defined relative to each other. These levels can cater to a cohort of students with wide-ranging levels of skills and abilities. As the course is more about basic programming and computational thinking skills, and less about using Python, some basic exercises can be made harder by disallowing certain Python
utilities such as sorting functions. We can also impose some constraints on the number of loops due to efficiency reasons, and as a result, the difficulty is increased. In the following three examples, we explain the difficulty levels for each exercise which has been used in our course. The solutions are not included here due to space constraints.

**Example 1**: "Prompt the user for a message. Display the message character by character on the same line but separate every two adjacent characters by a space". This is an "easy" exercise which tests basic string manipulation skills using a "for" loop in Python. Most students, including those who are new to programming, after being taught the concepts, can handle this exercise without issues. We have been observing that a few might need help with printing out all the characters on the same line. Note that the "easy" exercises are more than just trivial programming tasks such as creating a list, adding two numbers, etc.

**Example 2**: an "intermediate" exercise for string manipulation. The number of lines of code as well as the logic for a possible solution to this exercise is more complicated than that of Exercise 1. The student would have to figure out the exact range for a loop from $m$ to $n$, followed by a few complex conditional statements to get the correct output. Therefore, in the context of this paper, Exercise 2 is more difficult than Exercise 1.

**Example 3**: a "hard" exercise, also for string manipulation. This exercise is not as straightforward as Exercise 1 or 2, as it requires the student to keep track of the starting and ending point of each potential subsequence of letters using a Boolean variable, and the indices in the input. There is also the additional requirement of using a single "for" loop, which makes the exercise more challenging.

We note that it is not practical to manually create a lot of such ready-to-use exercises on-demand, i.e., when students need them, that are personalized to individual skill levels and learning concepts, as many students might be working concurrently on different kinds of problems at various difficulty levels. In the below section, we describe our approach to realizing automatic generation of ready-to-use exercises.

### 3. ExGen: Design and Implementation

Figure 1 shows the interactions between students, instructors and ExGen in our programming course. As ExGen aims to minimize instructor's involvement, the only role of the instructor is to provide a limited number of sample (seed) exercises with three difficulty levels: "easy", "intermediate", and "hard". Each student in the course can download these exercises into a personalized database which is saved locally. Students can load any exercises from their own

---

**Figure 1. Design and Implementation of ExGen**

---
database, work on them, and use ExGen to generate more exercises to practice. Students can choose to save new exercises into their personalized database for reference later.

ExGen provides functionalities for managing a personalized database of exercises for each student, generating new exercises via pre-trained LLMs, and providing a suite of auto-filtering checks to give students ready-to-use exercises at a difficulty level suitable for their knowledge and programming skills. In the below, we provide a brief description of each ExGen component as shown in Figure 1, and relevant implementation details.

3.1 Prompt management and LLM integration

ExGen generates new exercises by first constructing a "prompt" as a query to be sent to a pre-trained LLM such as OpenAI GPT-3.5-Turbo. It obtains a list of candidate exercises for further filtering before presenting a few selected ones in an IDE, e.g., VS Code, for the student to work on. Constructing good prompts for LLMs in many application domains, i.e., "prompt engineering", is currently an active area of research (Liu et al., 2023). LLM prompts may consist of questions, instructions, examples, etc., so that the model would reply with an output which has some desired qualities and/or quantities. In "zero-shot" prompting, we can query the model without providing any example of expected results. On the other hand, "few-shot" prompts provide several actual examples to the model.

In the GPT-3.5-Turbo model, which is used in ExGen's implementation, input prompts can be given as a series of messages with different parameters. Each message is an object with a role, which is either "system", "user", or "assistant", together with the content of the message. Usually, the first message would be a "system" message which helps define the behavior of the AI assistant. We can then alternate between "user" and "assistant" messages, in which "user" messages provide specific instructions for the AI assistant, and "assistant" messages can be used to provide examples of desired outputs, i.e., the "shots" in few-shot prompting. All messages must be constructed and passed to the LLM in a single API call, so that the model would have the right context when generating the new exercises. Note that OpenAI's web APIs do not have the memory of past requests. In this work, we consider two different strategies of prompting for exercise generation using GPT-3.5-Turbo, namely:

- **Zero-shot Prompting** (referred to as "zero-shot"): provides keywords on the programming concepts and difficulty levels required by a student, without using any examples of exercises. The prompt starts with a system message, e.g., "You are a helpful teaching assistant for undergraduates who are learning introductory programming in Python", and then adds a user message, e.g., "Give me three easy Python string exercises with this keyword: house".

  ![Figure 2. Messages for Few-shot Prompting](image)

- **Few-shot Prompting** (referred to as "few-shot"): provides a few actual examples on how the desired output should look like. In essence, when requested by students, ExGen constructs a series of messages following the GPT-3.5-Turbo model as shown Figure 2. ExGen first inspects the current exercise that the student is working on. Second, it searches for other exercises on the same concept, e.g., string, list, etc. and having a similar level of difficulty in the student's database. ExGen then sequentially appends the relevant examples, each with a problem statement, solution, and test cases, with the

---

Message 1: {"role": "system", "content": "You are a helpful teaching assistant for undergraduates who are learning introductory programming in Python."}
Message 2: {"role": "user", "content": "Give me a hard Python string exercise."}
Message 3: {"role": "assistant", "content": "Here is one hard Python string exercise: {example of a relevant exercise from the database}"}
Message 4: {"role": "user", "content": "Good. I want one more hard Python string exercise. Print the result with the same format as the previous ones."}
... (more messages for remaining examples)
Last message: {"role": "user", "content": "Good. I want {N} more hard Python string exercises using this keyword: {keyword}. Print the result with the same format as the previous ones."}
appropriate roles into the list of messages. The result will be like a conversation in which the LLM plays the role of a teaching assistant guided via a multi-turn conversation with several concrete examples. The last message is a user message requesting the LLM to generate N new exercises. Students can also specify some new keywords, e.g., "person", "movies", etc., so that the generated exercises would have more variations. Note that we can specify the difficulty levels, e.g., "easy", "hard", etc., and concepts, e.g., "string", "list", etc. in the "few-shot" prompt messages from Message 2 onwards in Figure 2.

In the implementation of "few-shot", ExGen uses three examples (3-shot) with the default parameters of the GPT-3.5-Turbo model. We deploy our own cloud instance for running LLM inferences using Azure OpenAI Service for better control and billing purposes. We also note that previous work in exercise generation, e.g., (Sarsa et al., 2022), mostly made use of the OpenAI Codex API, which has been deprecated since Mar 2023. The output of this stage is N candidate exercises, each having: 1) a problem statement, 2) a solution, and 3) test cases. ExGen will conduct auto-filtering for these exercises in the next stage.

3.2 Auto-filtering of generated exercises

In ExGen, we implement several methods to filter exercises generated by LLMs. The methods are chained, i.e., a newly generated exercise must successfully clear the previous filter before it is checked by the next filter, as shown in Figure 3. We consider standard software engineering techniques as well as LLM-based approaches for filtering. First, ExGen will check if the Python solution code can be compiled to bytecode without any errors. Then it will run all the unit test cases which are generated by the LLM together with every new exercise.

![Figure 3. Chain of Filters in ExGen](image)

After a new exercise can pass the first two filters, ExGen will do an additional check regarding the difficulty level of the exercise. The rationale for implementing this final check is that LLMs such as GPT-3.5-Turbo are probabilistic models, and their answers might be different from one interaction to the next. Therefore, it could be beneficial to ask an LLM to verify its own answers generated previously.

For the difficulty check, ExGen will perform a classification task by first constructing an appropriate prompt as shown in Figure 4, and then asking the LLM to classify the generated exercise in terms of difficulty levels. The prompt has two examples for each level of difficulty in a series of user and assistant messages. For the last user message, we ask the LLM to provide a classification for a new candidate exercise based on the provided examples. If the difficulty of the new exercise matches the required difficulty level, it will be shown to the student as a ready-to-use exercise using ExGen's VS Code extension. In addition, we observed that GPT-3.5-Turbo can generate exercises covering the concepts specified in the prompt such as lists, strings, etc. most of the time, so ExGen does not need to check if a requested programming concept is included in the output.

**Message 1:**

"role": "system", "content": "You are a classification model that will classify the difficulty of Python exercises."

**Message 2:**

"role": "user", "content": "I want you to classify this exercise: {example exercise}"}

**Message 3:**

"role": "assistant", "content": "Difficulty: " {example’s difficulty level}"

... (more messages for remaining examples)

**Last message:**

"role": "user", "content": "I want you to classify this exercise: {candidate exercise}"}
3.3 Coding interface and database management

ExGen provides a convenient interface for students to work and generate exercises right in VS Code. We have used ChatGPT as well as OpenAI's Chat Playground extensively, and we observed that it takes a lot of time and effort to copy/paste prompts and exercises for generation tasks. A screenshot of ExGen's interface is shown in Figure 5. Students can click on "+ More exercise" at the right bottom of the IDE to generate new ready-to-use exercises. They also have the option to change the problem statement, solutions or test cases used as examples in the prompt before clicking "Submit". ExGen will automatically connect to an LLM deployment, generate, and filter the exercises.

Students can also save selected ready-to-use exercises into his/her own personalized database. To make it easy for students, we use the TinyDB package, which is a portable Python document-oriented database, instead of setting up a fully functional SQL database engine like MySQL. The database comes bundled in ExGen so students can just use it right away without further setup or installation.

4. Evaluation

In this section, we aim to answer the following research questions (RQs):

- **RQ1**: Which is the best prompting strategy for generating ready-to-use exercises?
- **RQ2**: How effective is the auto-filtering approach implemented in ExGen?
- **RQ3**: How much time does it take to generate a ready-to-use exercise?

4.1 Dataset and performance measures

We used a seed set that has 49 manually crafted Python exercises including all levels of difficulty and introductory programming concepts such as string, list, dictionary, tuple, etc. There were 18 "easy", 17 "intermediate", and 14 "hard" exercises in the seed set. We then generated a total of 613 exercises (205 "easy", 270 "intermediate", and 138 "hard") using ExGen via the API of GPT-3.5-Turbo LLM. We then manually inspected each of these exercises to determine if they are ready-to-use. A ready-to-use exercise must satisfy the following criteria used in the manual inspection: 1) a clear problem definition, 2) a correct solution and test cases, and 3) a correct difficulty level as requested in the prompt.

We then measured the performance, i.e., accuracy, of each prompting strategy and filtering check when compared to the manual inspection results. For example, when "few-shot" prompting produces 5 ready-to-use out of 10 generated exercises, the performance of "few-shot" is calculated as 50%. On the other hand, the performance of a filtering check is calculated by how often it agrees with the manual inspection result. For example, the performance of the unit testing check is calculated as 70% if it provides the same result, i.e., ready-to-use or not, as the manual inspection for 7 out of 10 generated exercises. We measured the individual performance of each filtering check, as well as the overall filtering performance.
4.2 Results and discussion

**RQ1-Comparing different prompting strategies**: Figure 6(a) shows the percentages of ready-to-use exercises for the two prompting strategies and three difficulty levels. The "few-shot" prompting performed much better than "zero-shot" prompting in all cases. This is because "few-shot" incorporates examples to better guide the LLM in generating new exercises with appropriate levels of difficulty. We observed from the collected data that the "zero-shot" prompt usually produced many more trivial exercises, such as creating a list of several numbers, printing out a message with certain parameters, etc. These are not even at the "easy" level. This is understandable, as the "zero-shot" prompt does not have enough information about how an easy or hard exercise would be like. As a result, about 57% of "easy" exercises generated by "few-shot" prompting were ready-to-use, compared to just 31% of them in "zero-shot" prompting.

From Figure 6(a), it is observed that generating ready-to-use "intermediate" and "hard" exercises is quite challenging, even with recent advances in LLMs. In the experiments, only around 10% and 25% of "intermediate" exercises generated by "zero-shot" and "few-shot" prompting were considered ready-to-use, respectively. Similarly, we obtained only several ready-to-use "hard" exercises using "few-shot".

![Figure 6. Comparing two Prompting Strategies](image)

**Figure 6. Comparing two Prompting Strategies**

Figure 6(b) shows the percentages of generated exercises which have the correct difficulty levels (but might not be ready-to-use). By providing more detailed information about difficulty levels in the prompt, "few-shot" produced more exercises with the desired difficulty levels. Around 40% and 25% of exercises generated by "few-shot" could be considered as "intermediate" and "hard", respectively. However, quite a number of these exercises did not have clear problem statements, correct solutions and/or test cases, so they were not considered ready-to-use. Nonetheless, we observed that it's possible to make some small modifications so that many of these more challenging exercises would be ready-to-use.

<table>
<thead>
<tr>
<th></th>
<th>zero-shot (unit test)</th>
<th>zero-shot (difficulty)</th>
<th>zero-shot (chain)</th>
<th>few-shot (unit test)</th>
<th>few-shot (difficulty)</th>
<th>few-shot (chain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>68.6%</td>
<td>37.1%</td>
<td>69.5%</td>
<td>83.8%</td>
<td>96%</td>
<td><strong>83.8%</strong></td>
</tr>
<tr>
<td>Intermediate</td>
<td>51.4%</td>
<td>79%</td>
<td>90.6%</td>
<td>55.3%</td>
<td>78.8%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Hard</td>
<td>51.5%</td>
<td>90.9%</td>
<td><strong>98.5%</strong></td>
<td>54.2%</td>
<td>77.8%</td>
<td><strong>97.2%</strong></td>
</tr>
</tbody>
</table>

**RQ2-Evaluating ExGen’s auto-filtering approach**: The standalone performance of the unit testing and difficulty checks, as well as the overall filtering performance (chain of all checks) are shown in Table 1 (higher percentage is better). From the experiments, we noted that the solutions of most generated exercises could pass the compilation check so that its result is not included in Table 1. The unit testing check when used alone could correctly identify if a
candidate exercise is ready-to-use or not in many cases, especially for "easy" exercises generated with "few-shot" prompting (83.8% accuracy when compared to manual inspection's result). However, unit testing became less effective as the only filter for harder exercises as shown in Table 1. This is because although around 50-70% of candidates for harder exercises had working solutions and test cases in our experiments, they were not at the right levels of difficulty, i.e., "intermediate" or "hard".

Table 1 also shows the performance of the difficulty check when used alone as an exercise filter. It is not very effective when filtering "easy" exercises generated by "zero-shot" prompting as there could be many trivial exercises. The difficulty check performed better when filtering "intermediate" and "hard" exercises, compared to the unit testing check.

We observed good performance for our auto-filtering approach when chaining all the three filters, i.e., compilation followed by unit testing followed by difficulty check. The accuracy of the chained filters improved significantly for all prompting strategies when filtering "intermediate" and "hard" exercises. Overall, we noted that "few-shot" combined with the chained filters produced the best performance. Using this combination, we are more confident that ExGen can provide ready-to-use exercises for students to work on right away.

It was also noted that the LLM used in ExGen tends to provide partially correct solutions and/or test cases when generating "intermediate" and "hard" exercises. Although these candidate exercises did meet the difficulty requirement, they failed ExGen's chained filters in the end. We believe that with a few manual modifications by human experts, they can be ready-to-use by students.

**RQ3-Exercise generation time:** We measured the average time to produce a ready-to-use exercise using the two prompting strategies combined with the chained filters for various difficulty levels. We noted that the generation time was quite reasonable for "easy" exercises, which was around 10 and 20 seconds per exercise with "few-shot" and "zero-shot" prompting, respectively. An "intermediate" exercise required 53 seconds for "few-shot", and 90 seconds for "zero-shot". Using a less effective prompt like "zero-shot" increased the average generation time as there would be more candidate exercises which were not ready-to-use.

On the other hand, generating ready-to-use "hard" exercises could take too long to be considered on-demand (about 8 minutes for "few-shot", and 14 minutes for "zero-shot"). This was because many candidate exercises would have to be generated before ExGen could find one that met the "hard" difficulty requirement and has correct solutions/test cases. To this end, we believe that "hard" exercise generation should be best done in a hybrid way. That is, ExGen can provide initial generation and filtering support for course instructors to quickly modify the LLM's output and make it ready-to-use.

**Discussion:** From the careful evaluation using more than 600 auto-generated exercises, we believe that ExGen can provide students with ready-to-use programming exercises customized at the right difficulty levels. LLM prompting strategies having sufficient details such as examples have been shown to generate more exercises that are ready-to-use. In most cases, our auto-filtering approach with a chain of software engineering and LLM-based checks could work well regardless of the prompting strategy used. However, it is noted that automatically generating truly ready-to-use, "hard" exercises is still a problem in AI-enabled education despite recent progress in LLM and AI development.

**Threats to Validity:** We note that there are several limitations which may affect the validity of this study. First, the output from LLMs could be non-deterministic, which might impact the quality of generated exercises. We have tried to take this into account by evaluating many exercises. Second, manual evaluation of exercises by human experts could be subjective, e.g., when assigning a difficulty level. For this, we have used the opinion of two different programming experts on the set of generated exercises. Finally, as more advanced LLMs such as GPT-4 API are still in limited availability, it is possible that the results obtained from these models are different from what we have seen in this study. We plan to do further investigations using exercises in Python and other programming languages once access to more powerful LLMs becomes widely available.
5. Related Work

Much research has been done in automatic generation of formative feedback and reference solutions to code produced by students (Keuning et al. 2018, Koutcheme 2022, Ta et al. 2022). Such feedback and reference solutions could be generated by current pre-trained LLMs to help novice programmers know how to proceed when facing coding issues. On the other hand, not much has been done in generating ready-to-use programming exercises for students to do more practices. Kurdi et al. (2020) conducted a systematic review of automatic exercise generation in many different domains such as analytical reasoning, geometry, history, logic, relational databases, programming, and science.

Zavala and Mendoza (2018) used Automatic Item Generation (AIG) to address the problem of creating many similar programming exercises using pre-defined templates which are used for quizzes. The main goal was to ensure consistency in testing many students with questions of the same level of difficulty. On the other hand, ExGen focuses on generating ready-to-use exercises for a specific difficulty level and concept that the student is currently working on. We leveraged the latest advances in LLMs to autogenerate many novel exercises and filter them to ensure that they are suitable for students. Exercises considered in ExGen are different from other kinds of programming practices such as faded Parson problems (Fromont et al. 2023), which require students to fill in code in partially scrambled solutions.

The most relevant work in automatic exercise generation using pre-trained LLMs has been done by Sarsa et al. (2022). The authors explored OpenAI Codex (which has been deprecated since Mar 2023) for the purpose of creating new programming exercises and code explanations. They found that many Codex-generated exercises were sensible and novel, but may have confusing problem statements, missing or faulty test cases. Sarsa et al. (2022) did not consider the implementation of an auto-filtering tool to supply students with ready-to-use exercises at various difficulty levels.

Recently, there has been an increase in the number of works leveraging pre-trained LLMs for educational purposes. Kasneci et al. (2023) discussed the potential benefits, for instance content generation and personalized learning, as well as challenges, e.g., model biases, system brittleness, etc., when applying LLMs to education. Similarly, Becker et al. (2023) elaborated the educational opportunities of AI code generation, and how educators should act quickly given these developments. Finnie-Ansley et al. (2023) reported the performance of OpenAI Codex on real questions from programming courses when compared to that of actual students. Denny et al. (2023) studied GitHub Copilot, a plug-in for IDEs like VS Code, which is based on the Codex model, to see what kinds of problem it would not perform well. The authors also found that prompt engineering can play an important role in interacting with AI tools like Copilot. MacNeil et al. (2023) found that code explanations such as line-by-line, high-level summary, and lists of important concepts generated by Codex and GPT-3 were helpful to most students. Kazemitabaar et al. (2023) studied how Codex's code generation capability could assist novice programmers via a controlled experiment for young students. We note that our work focuses on automatic exercise generation and filtering, not feedback and explanation for code submitted by students.

6. Conclusion

We have implemented and evaluated ExGen, a new software tool that automatically generates ready-to-use exercises for computing students. ExGen manages a personalized database of seed exercises and constructs appropriate LLM prompts to obtain candidate exercises for a requested difficulty level and learning concept. ExGen incorporates a novel combination of auto-filtering checks which reduce the tedious work of manually inspecting all the generated output. Our extensive evaluation with more than 600 programming exercises generated using GPT-3.5-Turbo API demonstrated that ExGen can be used to provide students with ready-to-use exercises. We plan to investigate new approaches for automatic prompt optimization so that harder exercises can be generated more easily. As a VS Code extension, ExGen can be released to our incoming freshmen for further user evaluation in the next academic year.
Acknowledgements

This research is supported by the Ministry of Education, Singapore, under its Tertiary Education Research Fund (Award No. MOE2021-TRF-014). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References


Object Identification Training Support System for Object-Oriented Design with Cooking Recipes
Daiki Maeda\textsuperscript{a}, Kota Kunori\textsuperscript{b} & Tomoko Kojiri\textsuperscript{b}
\textsuperscript{a} Graduate School of Science and Engineering, Kansai University, Japan
\textsuperscript{b} Faculty of Engineering Science, Kansai University, Japan
* k554875@kansai-u.ac.jp

Abstract: In object-oriented programming (OOP), each process is accomplished by the interaction of objects. In the first step of designing a system using OOP, programmers need to decompose a given use case into smaller sub-processes and identify objects. However, for novice programmers, it can be difficult to imagine things that are not explicitly stated in the use case, and some may not understand that the process is achieved by the interaction of objects. We have therefore developed a training method that identifies objects by connecting them with things that novice programmers can more easily imagine, such as recipes. Similar to use cases, recipes are a set of processes whose subjects are not clearly mentioned. When we decompose a recipe set into smaller sub-processes made up of kitchenware items, it is possible to recognize these items when imagining the cooking scene. We also developed a system that helps novice programmers to extract the appropriate kitchenware from recipes. Through the experience of rewriting a recipe utilizing the process of cooking, novice programmers are able to recognize that there are subjects hiding in a process description whose subjects are not mentioned, which helps them understand how to identify the hiding objects.

Keywords: object-oriented programming, extraction of objects, recipe, training support system, drawing scene

1. Introduction

Object-oriented programming (OOP) is widely utilized in the software industry (Teif & Hazzan, 2006). In OOP, program behavior is implemented as an interaction between objects. When designing software using OOP, the classes of objects need to be designed in accordance with the requirement specifications (e.g., use case). To design a system so that the degree of coupling between classes is low and the degree of condensation within classes is high, it is necessary to identify functions that can be grouped together and to identify the things that can carry those functions as objects. However, it can be difficult for novice programmers to identify objects from the use case that are not explicitly mentioned, and some do not understand that the process is accomplished by the interaction of objects. Therefore, in extreme cases, they may design a system with a single object called “system” and assign all functions to it.

Many studies have been conducted to help programmers understand the object-oriented concept (Lian, Varoy, & Giacama, 2022; Seng, Yatim, & Hoe, 2018; Dwarka & De Villiers, 2015), some of which specifically focus on the behaviors of the objects (Abidin & Zawai, 2020; Gestwicki & Jayaraman, 2005). Although these studies provide an understanding of the object-oriented concept, they do not support the design process of OOP, which is essentially to identify objects in accordance with the requirement specifications. In order to identify objects, it is necessary to imagine the potential interactions of multiple objects from the given requirement specification.

In the current study, we explore two key research questions:
1) What kind of thinking is necessary to derive adequate objects from the process description in which subjects are not clearly mentioned?
2) What kind of training is appropriate to foster such thinking?

To derive the answers to these questions, we developed a training method and evaluated its effectiveness.
The proposed training method identifies objects by connecting them with things that novice programmers can more easily imagine, such as recipes. Similar to use cases recipes are a set of actions whose subjects are not clearly mentioned. Our method circumvents this difficulty by decomposing the set into smaller sub-processes of kitchenware. For example, “to cook meat” can be rewritten as “a stove heats up a frying pan and the frying pan cooks the meat.” We can recognize the kitchenware items when we imagine the cooking scene, so it is easier to grasp that the recipes are essentially a series of kitchenware interactions and thereby to imagine what kind of kitchenware is involved in the recipe. Our system basically encourages novice programmers to identify kitchenware from the written recipe that they imagine from its scene. The system asks the users to 1) extract the kitchenware from the recipe, 2) rewrite the recipe in accordance with the extracted kitchenware, and 3) draw a picture using the identified kitchenware. This process is repeated until they are able to accurately depict the cooking scene that they imagine from the recipe. Through the experience of rewriting a recipe utilizing the process of kitchenware, novice programmers are able to recognize that there are subjects hiding in the process description whose subjects are not mentioned. In addition, by continuing this activity until they draw the scene that they imagine, they begin to grasp that imagining the scene is necessary for deriving adequate objects.

2. Approach

2.1 Requirement Specification of System Development

A requirement specification is a document that summarizes the functions and characteristics that a system to be developed should have. It consists of a base action series, exemplified action series, and other conditions, and it is established in the requirements definition phase of system development to form the policy for the subsequent development plan. An example of a requirement specification is a use case description, which describes specific interactions between an actor and a system in that scene, as well as related conditions (Technologic Arts Inc., 2009). Since actors are the users of the system, the external hardware, the external systems related to the target system, etc., the use case description is a summary of the requirements from the user's point of view. Therefore, the configuration of the system (e.g., the objects) are not typically mentioned.

Figure 1 shows an example of the basic action series for the use case “withdrawing money from an ATM”.

| 1. Select “cash withdrawal” |
| 2. Enter your PIN |
| 3. Enter the amount of money |
| 4. Confirm the amount displayed on the screen and press the confirm button |
| 5. Take out the amount of money you wish to withdraw from the ATM withdrawal slot |
| 6. Receive cash from the withdrawal slot |

*Figure 1. Basic action series for cash withdrawal from an ATM.*

2.2 Object Identification Method

Since the basic action series of a use case description does not always mention the subject or target of each action, the programmer will need to derive objects that can be the subjects or targets. If the scene in the use case can be imagined, the identified objects should be in the imagination.

We propose a method for deriving adequate objects by having the programmers draw a picture of the processing scene. Since the adequate objects are those described in the picture, this method encourages programmers to derive subjects and targets from the use case that are the same as what they imagine. By having them draw the picture, they become aware that the objects they add to the image are the objects that they should identify.
2.3 Acquisition of Object Identification Methods Using Recipes

In order to acquire the thinking method that the objects in the scene of the use case specification are the objects to be identified, it is desirable to train the thinking method by a use case whose scene is easy to imagine. In this study, we utilize a cooking recipe as the substitution for the use case.

A recipe is a sequence of actions to take when preparing a specific dish. In this context, the subjects of the actions can be regarded as the kitchenware items, and some data (e.g., ingredients) can be viewed as objects being processed by the kitchenware. If we view an item of kitchenware as a subjective object, the recipe can be viewed as the interaction of multiple kitchenware items. However, in a typical recipe, the kitchenware is not described clearly, if at all. In this respect, we believe that recipes are similar to use cases. If it becomes possible to identify kitchenware from recipes, novice programmers can learn how to identify objects from the use case and thereby successfully identify objects. This is why our method utilizes a recipe, not a software use case, as the training target. We came up with the following steps to train the thinking of object identification.

Step 1. Derive the objects (e.g., subject and target) corresponding to each action of the recipe.
Step 2. Rewrite the action with the derived objects.
Step 3. Draw a picture of each cooking action in the recipe using the identified objects.
Step 4. Check whether the picture in Step 3 satisfies the imagined scene. If not, identify the missing objects and return to Step 2.

As an example of the steps, let’s focus on the cooking action “boil potatoes” in Figure 2. If we assume the subject of this process is a “person” and the target is a “potato” (Step 1), the cooking action can be rewritten as “a person boils potatoes” (Step 2). If we draw a picture like the one in Fig. 3 using the extracted objects (Step 3) and find that it is different from the imagined scene, as shown in Fig. 4, we extract the missing objects by examining the differences between the drawn picture (Fig. 3) and our imagination of the scene (Step 4), such as “stove”, “fire”, “pot”, and “water”. We then return to Step 2 and rewrite the “boil potatoes” action on the basis of these derived objects as “1. a person turns on a stove, 2. the stove turns on the flame, 3. the flame heats a pot, 4. the pot heats the water, and 5. the (heated) water boils the potatoes”. In Step 3, we draw the scenes from these objects and, if we satisfied with the picture, the process ends, indicating that we have derived adequate objects. If not, we go back to Step 4 and look for the missing objects again.

1. Boil potatoes
2. Mash potatoes
3. Cut onions
4. Cut carrots
5. Heat onions
6. Heat carrots
7. Cut cucumbers
8. Cut ham
9. Mix mayonnaise

Figure 2. Potato salad recipe.
Figure 3. Cooking scene drawn in accordance with identified objects.
Figure 4. Imagined cooking scene.

2.4 Overview of Object Identification Training System

The object identification training system is intended to provide a place where users can experience how to identify objects by using recipes, thereby enabling them to acquire the thinking necessary to identify objects from the process. A system configuration diagram is shown in Fig. 5. The system consists of two interfaces and a feedback function, and it utilizes recipes in a recipe database (DB) for the training.
The identification interface is a screen for inputting the objects identified from each action of the recipe. The user inputs the subject and target from the displayed recipe sentence and uses them to rewrite the sentence. The confirmation interface is a screen where the user draws a picture using the objects (subject and target) input to the identification interface to confirm whether adequate objects have been input. After the picture is drawn, the confirmation interface gives a comment to the user asking if there are any lacking objects. If not, it finishes the process. If yes, it activates the identification interface again. The identification interface asks the user to rewrite the recipe sentence using the lacking objects. It then checks whether the written sentences represent the action of the original process, and if not, it gives feedback to modify the sentences again.

3. Feedback Function

A sentence corresponding to an action in the recipe represents that the subject executes the action and the target is manipulated by the action. If the sentence is rewritten by using several sentences related to sub-actions, it is important that the first subject of the sub-actions, the last target of the sub-actions, and the last action be the same as the original action. The subjects and targets of other actions are intermediary; that is, instead of indicating a direct manipulation of the target by the subject, it means that others are told to manipulate the target. Figure 6 illustrates the relations between the original action and the sub-actions. Suppose that the original action P is a process from subject A to target B. If the action is replaced by sub-actions, the first subject needs to be A, the last target needs to be B, and the last action must be P. Other actions (a, b, and c) and other objects (X, Y, and Z) might be placed at various positions between them.

The feedback function checks whether the rewritten recipes satisfy such conditions. In addition to the first subject, the last target, and the last action, it needs to check that there are no omissions in the transfer of actions. The actions are connected when the target of the former action and the subject of the latter action are the same. By considering these conditions, the feedback function determines whether the rewritten recipe is the same as the original one by the following steps:

Step 1. Determine the subject and the target of each sub-action in the rewritten recipe.
Step 2. If the previous target and the subsequent subject are the same, eliminate the target/subject and previous action and combine them into one action.
Step 3. If the rewritten recipe is changed to one action, compare its subject, target, and action to the original action.

Here, we show an example with the original recipe “a person boils potatoes” and the rewritten recipe as “1. a person starts up a stove, 2. the stove turns on a flame, 3. the flame heats a pot, and 4. the pot boils the potatoes”. Since the target of 1 and the subject of 2 are the same (i.e., “a stove”), these actions are combined as “a person turns on the fire of a stove”. As a result, we have “1. a person turns on the flame, 2. the flame heats a pot, and 3. the pot boils the potatoes”. Since the target of 1 and the subject of 2 are the same in the rewritten recipe, they are combined to create “a person heats a pot”. As a result, “1. a person heats a pot, and 2. a pot boils the potatoes” is created. In the same way, “1. a person boils the potatoes” is generated. Since it is the same as the original recipe, the rewritten recipe is regarded as the same action as the original one.
The feedback function provides feedback in accordance with the result. If the original recipe and the rewritten one are the same, it gives “The re-written recipe is appropriate.” If not, it points out the inappropriateness and suggests how to fix it. There are two possible causes of inappropriateness. The first is that there is an omission in the transfer of processing. For example, if the sub-actions are “1. a person starts up a stove, and 2. a pot boils the potatoes,” some actions might be missing between the two actions, since the target of action 1 and the subject of action 2 are different. In this case, the feedback function shows the user where the cooking steps are not connected and gives a hint if the subject and target are different. The second cause is that the combined action is different from the original one. In this case, the system gives the comment “The action in the original recipe and the sub-actions in the rewritten recipe are not the same”.

4. Prototype System

4.1 Identification Interface

Figure 7 shows an image of the identification interface screen. When a recipe name is selected, the actions in the selected recipe are displayed in the recipe display area. The user needs to input the subject, target, and action extracted from the recipe into the cooking action input area. By selecting a row in the cooking action input area and clicking the “Add Row” button, a new row is added under the selected row. By clicking on each cell of the row, the subject, target, and process can be entered. To delete the row, click on the row to be deleted and select the “Delete” button. The order of the input in the cooking action input area represents the order of the cooking from start to finish. When the “Judgment” button is clicked, the feedback function evaluates whether the input cooking actions are appropriate and displays the result in a message box.

![Figure 7. Identification interface.](image)

4.2 Confirmation Interface

The confirmation interface is a screen for creating a picture of the cooking scene for each action using the objects identified in the identification interface. Figure 8 shows an example of the confirmation interface screen. The user can select the action to be depicted in the cooking action display list. When a target action is changed, a new canvas for the action appears in the picture creation area, and the user can draw a picture representing the scene of the action. The picture is drawn by placing the images displayed in the image display area onto the picture creation area. The image display area displays images of objects stored in the system in advance that may appear in the cooking scene. The user can also add their own images by dragging and dropping image files. To draw the picture, a user selects an object from the list of identified objects, clicks on the image representing the object, and then clicks on the “Match
String to Image” button to display the corresponding image. The user can also move, delete, and change the size of the image. When the cooking scenes of all actions are drawn using the identified objects and the “Finish Creation” button is clicked, the system presents a message to consider whether the drawn pictures represent the user’s imagination of the scene. The intent of this message is to prompt the user to think about whether there are any missing objects. If there are, the system moves to the identification interface again and lets the user subdivide the cooking action using the missing objects recognized in the confirmation interface. If there are no missing objects, the operation terminates.

5. Conclusion

We have developed an OOP training system for identifying objects that are responsible for the sub-actions of requirement specifications, such as use case descriptions. The training system introduces a recipe instead of a use case as a training target, which makes it easier for novices to imagine the objects and their interactions in the scene. Our system helps the user to identify valid objects by having them draw a picture of the cooking scene with the identified objects. It also provides a feedback function to check whether a rewritten recipe using the identified objects represents the same actions as the original recipe.

The recipes handled by the system in this study were linearly connected, with only one subject and one target per process. However, there are also actions that have multiple targets. In such cases, some actions are performed in parallel and are therefore not connected linearly. To cope with such recipes, the feedback function needs to be improved so as to check the validity with not only simultaneous actions but also actions occurring separately.

References


Development of Learning Support System for Critical Reading of Academic Papers

Aota NISHIDA\(^a\), Kazuhisa SETA\(^b\) & Yuki HAYASHI\(^b\)

\(^a\)College of Sustainable System Sciences, Osaka Prefecture University, Japan
\(^b\)Graduate School of Informatics, Osaka Metropolitan University, Japan

*seta@omu.ac.jp

Abstract: Critical reading skills, which require learners to examine texts from multiple perspectives rather than relying solely on what is written, play an essential role in reading academic articles to acquire high-level literacy. Several systems have been developed to support the improvement of critical reading skills; however, there is no system that follows the learners' thinking and the process of reading articles at the content level and provides feedback that captures the semantic connections specific to the research domain. Therefore, we propose a support system for critical reading of academic papers, which has a function to help learners organize the contents of academic papers to promote awareness of essential viewpoints that encourages their critical reading based on semantic connections specific to the research domain.

Keywords: Critical reading of academic papers, research activity ontology, dynamic intervention based on ontology

1. Introduction

Reading and writing skills such as understanding advanced texts, critically examining them, and writing papers are called "high-level literacy," the need to develop such skills has been pointed out in recent years in secondary and higher education. In response to the expansion of the number of learners who enter universities and to social demands, many universities have begun offering various study and academic skills as part of first-year education or remedial education. Critical reading skills, which require learners to examine texts from multiple perspectives rather than relying solely on what is written, play an essential role in reading academic articles to acquire high-level literacy. To develop these skills, it is essential to understand the writer's thoughts in a way that allows the reader to follow their own thoughts and form their own ideas after thoroughly examining them (Norris, et al., 2013, Duke, N. K., & Carlisle, J. F., 2011, Pearson, P. D. et al., 1992).

Okochi states that there are three aspects to critical reading activities (Okochi, 2012) (Figure 1). The first is the aspect of "examining one's own understanding" (comprehension monitoring) by examining whether one understands the author's argument correctly, the second is "examining the writer's logic" by examining the logical line of the writer, and the third is the productive aspect of "deepening one's thinking" by considering new problems and alternative solutions to the writer's logic that one has understood.

In general, we define "critical reading" as productive activities to deepen one's own thinking while understanding and examining the author's logic while reading academic articles, and this study aims to develop critical reading skills.

In this study, we encourage learner's self-dialogue in which the author's thoughts and one's own thoughts on the author's thoughts formed from the author's argumentation are interacted. Based on the idea, we propose a support system for critical reading of academic papers, which has a function to help learners organize the contents of academic papers by comparing their ideas with those of authors and a function to compare the organized contents of reading and to promote awareness of critical viewpoints.
2. Critical Reading Support System

We developed a critical reading support system that encourages productive dialogue. The system interface is shown in Figure 1. The system has a function to support the organization of the contents of an article by comparing the author’s and one's ideas and a function to compare the organized reading contents and promote awareness of critical perspectives.

The system is a server-client model that can be used anywhere and anytime. By uploading a PDF file of an article, learners can read it in the same format as they usually read it and comprehend the article with their dialogue. The learner reads the paper displayed in the paper display area (Fig. 1(ii)). In the question presentation area (Fig. 1(iii)), a question that encourages deepening and consideration of the content of the paper (e.g., "What are the difficulties that learners have that prevent them from achieving their learning goals?") is presented by the system. Using this question as a guide for reading comprehension, the Thought Organization Map (Fig. 1(i)) provides a thinking organization environment that allows the learner to separate the author's ideas from his or her own and to organize them relative to each other (Mori, N., et al., 2020). The learner reads the article while marking it and expresses his or her understanding of the author's arguments and ideas on the map using the question.

2.1 Thinking Organization Map

When a learner adds a question that he/she thinks should be considered as a question node on the thought organization map, a purple "prediction node" labeled "your prediction" is automatically added as a child node to encourage the formation of his/her idea (Fig. 2(i)). When the learner responds to this prediction node, an orange "interpretation node" labeled
"author's claim" is automatically added (Fig. 2(ii)). As the learner continues to read, he or she is encouraged to dig deeper relative to his or her prediction, editing and positioning the node's contents when the author's claim appears later. A "critique node" that critiques the author's ideas, labeled "Your Critique," is also automatically added (Fig. 2(iii)). This "critique node" requires the assignment of semantic tags that express the critique viewpoints (value judgments, issues, opinions, etc.), which are hierarchically arranged in advance, so that learners can compare their predictions with the author's claims and deepen their thinking. In this way, the nodes are presented step by step in the order of "prediction," "interpretation," and "critique" for each question, and these cognitive activities are aimed at making the author's claims relative to their own. These are performed in a chain-like manner, systematically representing the process of internal self-dialogue during reading in a mind map format.

2.2 Cross-referencing Function

Cross-referencing function between a node and article texts is also provided. A node and the text it refers to are highlighted in the same color according to the node type. This is achieved by converting the article file read in pdf format to HTML format, enclosing each text of the article with tags and giving it a unique ID, and assigning the unique ID of each node to the text information selected by the learner. This mechanism allows the system to capture the thoughts of multiple learners who have read the same paper and what they read. Thus, the system enables learners to organize their thinking while examining the contents of the paper and the results of their thinking as appropriate during reading and to compare the reference structure of the paper with the maps of other learners who have read the same paper after reading it.

3. Concluding Remarks

This study focused on three aspects of developing critical reading skills. We defined critical reading as a productive activity that deepens one's thinking while understanding and examining the author's logic in reading academic articles. The functionality of the system could only be realized based on the ontology in which concept definitions specify semantic relationships specific to the intelligent tutoring system research domain (Chen, L., & de Oliveira, R., 2021, Mizoguchi, R., 2003).

We will verify whether the critical reading comprehension support system described in Section 2 can be used in actual practice. After confirming that the system works as intended, we will conduct a questionnaire survey to determine whether using the system to read papers has a potential leading to a change in learners' attitudes toward critical reading comprehension.

References

Overcoming Barriers to Sustainable Dissemination of L2 Learning Resources: An Integrated Framework for Creating and Distributing Dialogue Scenarios

Emmanuel AYEDOUNa*, Yuki HAYASHIB & Kazuhisa SETAb

aFaculty of Engineering Science, Kansai University, Japan
bGraduate School of Informatics, Osaka Metropolitan University, Japan
*emay@kansai-u.ac.jp

Abstract: Language learning resources, such as dialogue scenarios, are important for supporting second language (L2) learning by engaging learners in simulated conversations or role-plays. However, there are often barriers to the wider dissemination of these resources, including the lack of a standard format for distributing them, the difficulties of maintaining and updating them over time, and the need for effective means of tracking their use and impact. This paper presents an integrated framework for fostering the wider dissemination of L2 learning resources, with a focus on dialogue scenarios. The framework is designed to address the above challenges by providing a novel approach for creating and distributing dialogue scenarios, as well as a set of guidelines to ensure that they are pedagogically sound and compatible with learning management systems (LMS). The paper also discusses the potential benefits of the framework for L2 learners and educators and outlines future directions for the project.

Keywords: Dialogue Scenarios, L2 learning, Authoring interface, SCORM-compliance

1. Introduction

The Second language (L2) learning has the potential to be a transformative experience, opening up new opportunities for communication and cultural exchange. Yet, many learners are unable to access high-quality language learning resources due to financial, geographic, or technological barriers (Zimmerman, 2014).

On the other hand, the use of learning management systems (LMS) in education, and more specifically in computer-assisted language learning, has become increasingly common in recent years, as they provide a convenient platform for organizing and delivering course materials, managing assessments, gathering learning data, and facilitating communication between learners and instructors (Watson & Campana, 2017). In order for an LMS to be effective, it must be properly configured and populated with high-quality learning resources (Kataoka et al., 2018).

For instance, learning resources such as dialogue scenarios are an effective tool for L2 learning, as they can engage learners in simulated conversations or role-plays, helping them to develop their language skills and confidence in using the target language (Willis & Willis, 2007). Overall, the use of dialogue scenarios in L2 learning can support learners’ communicative competence (Johnson & Valente, 2008), cultural awareness, and can contribute to a more effective and engaging language learning experience (Ayedoun et al., 2021). However, the distribution of these learning resources over LMS has historically been limited, with most being developed for a specific course or program and not widely available for reuse. In addition, the implementation of new dialogue scenarios requires a significant amount of knowledge engineering effort and expertise in dialogue design (Ayedoun et al., 2021), which can be a barrier to their wider dissemination in learning management systems.
Other barriers may include the lack of a standard format for distributing resources, the difficulties of maintaining and updating materials over time, and the need for effective means of tracking their use and impact.

In this paper, we present an integrated framework for overcoming these barriers and fostering the sustainable dissemination of L2 learning resources, with a focus on dialogue scenarios. The framework is designed to provide a standard format for creating and distributing resources, as well as a set of guidelines to ensure that they are pedagogically sound and compatible with learning management systems (LMS). Then, we discuss the potential benefits of the framework for L2 learners and educators and outlines future directions for the project.

2. Proposed Framework

We identified several key requirements for building a desirable Integrated Language Learning Support Framework (I2LSF) that puts educators at the heart of learning resources design for a more sustainable dissemination of L2 learning resources. We propose that such an integrated framework should ideally be flexible, collaborative, easy to use, analytics-driven, integrate with other tools, and inclusive. Additionally, it should have a sustainable plan, and focus on accessibility to all learners.

Moreover, we also anticipate that such a framework could lead to the development of cost-efficient intelligent tutoring systems (ITS) that have the ability to reach a large number of students while still maintaining a high level of personalization, adaptivity, and interactivity. In other terms, as ITS are scaled up and move beyond the standard three or four-component model, there is a potential for adaptation to users to occur at various levels of granularity. Previously, adaptation in tutoring systems was thought to consist of two main components: “macro-adaptivity” or the “outer loop”, and “step-based adaptivity” or the “inner loop” (VanLehn, 2006). However, the I2LSF framework as proposed here, and fulfilling the prerequisites described above may also hold the potential for “meta-adaptivity” in which the system adapts to the user by directing them to a different language learning activity that may suit better learners’ needs. Additionally, research on “micro-adaptivity” has explored the benefits of using learning data to fine-tune interactions at a level below the problem step, such as keystroke-level inputs, emotion detection, and the presentation and timing of feedback (Graesser, 2011). In the context of the I2LSF framework, we believe that micro-adaptivity may be achieved at the interaction level with the presentation of specific conversational feedbacks.

In sum, from the above and inspired by Nye’s previous work on the potential of intelligent support systems to improve education in resource-constrained countries (Nye, 2015), we propose that a desirable learning support framework for creating and distributing learning resources in the context of L2 learning should feature the following four connected

![Figure 1. Outlines of our proposed I2LSF with the four levels of adaptation.](image)
levels or loops of adaptivity: macro-adaptivity, step-based adaptivity, meta-adaptivity, and micro-adaptivity. These four loops are likely to remain important decision points for designers in the I2LSF, as they can be thought of as boundaries within the learning support system. In addition, each of these types of adaptivity has its own advantages and can be used to provide a more effective and efficient learning experience, by adjusting to the student’s needs, preferences, and progress. Such an I2LSF could also be helpful to teachers in teaching language by providing them analytics on how the students are progressing and giving them suggestions on how to improve their teaching. On the other hand, incorporation of AI based techniques such as Adaptive learning, Self-generated data and Natural Language Processing can make the system more effective and efficient in the way it assesses the learner’s progress and gives the feedback which can help the learner to identify his strong and weak areas.

In sum, from a theoretical point of view, if achieved, the I2LSF framework proposed in the context of study hold the potential to allow educators to deliver learning resources that can be adjusted and tailored to the individual needs and preferences of each student with user adaptation being carried out at four levels, as depicted in Figure 1.

3. Potential and Future Directions

While our project is still in the early stages of development, we believe it has the potential to make a significant impact on the field of L2 learning and are excited to see how it evolves in the future. In addition to providing access to a larger number of high-quality resources, the framework has the potential to support the professional development of scenario authors by giving them a platform to share their work and receive feedback from their peers. Another advantage of the framework is that it makes learning resources available from anywhere with an internet connection, making language learning more convenient and accessible. It could also facilitate collaboration among authors and institutions, enabling the creation of more complex and nuanced scenarios that better reflect the diversity of L2 learners.

To evaluate the effectiveness of the proposed framework, we plan to conduct a series of case studies with a diverse group of L2 learners and educators. These experimental evaluations will examine the impact of the framework on learners’ language skills and attitudes towards L2 learning, as well as the adoption and use of the platform by scenario authors.

References


Learning Support System to Understand Others Through Dramatic Script Reading and Its Evaluation

Hanano OKAMOTO\textsuperscript{a}, Yuki HAYASHI\textsuperscript{b} & Kazuhisa SETA\textsuperscript{b}

\textsuperscript{a}College of Sustainable System Science, Osaka Prefecture University, Japan
\textsuperscript{b}Graduate School of Informatics, Osaka Metropolitan University, Japan
\textsuperscript{*}hayapy@omu.ac.jp

Abstract: To form and maintain smooth interpersonal relationships in various situations in social life, the ability to appropriately understand the mental states of others, such as their intentions and feelings, is required. However, it is difficult to infer the unobservable mental states of others, and it is also difficult to learn from others how to perform inferring activities that are not externalized. In this study, we focus on “dramatic scripts” as a learning material that contributes to the training of the ability to understand others and propose a learning support system that aims to improve the ability to understand others and to cultivate an attitude of understanding others through inference activities about the emotions and conflicts of characters.

Keywords: Ability to understand others, dramatic scripts reading, learning support system

1. Introduction

It is required to form and maintain smooth interpersonal relationships in various situations in social life. For this purpose, it is necessary to appropriately understand the mental states of others, such as their intentions, beliefs, and feelings (Apperly, 2012; Epley & Caruso, 2012). An increased understanding of the mental states of others will lead to consideration for others and two-way communication, thereby promoting smooth interpersonal relationships. In addition, an attitude that strives to understand others who have different backgrounds from oneself is an important element forming the basis for understanding others (Shih et al., 2009; Stephan & Finlay, 1999).

On the other hand, it is not always easy to infer the mental states of others that cannot be explicitly captured because the intentions and emotions of others are not always expressed in everyday communication. When the actual feelings of others differ from their expressed behaviors, it is necessary to infer their emotions based on contextual factors other than facial expressions, such as the background situation in which they are placed. Since the mental states behind such utterances and actions are complex and chaotic (Berrios, 2019), there is no unique model for inference activities, and it is also difficult to learn from others such inference activities inside an individual.

In this study, we focus on the “dramatic scripts” as promising learning material for cultivating the skill of understanding others. In dramatic script reading, it is required to understand the mental states of others from the viewpoint of characters with different backgrounds from the reader by inferring the intentions and feelings of the characters based on the clues in the chain of utterances (dialogues) made by the characters. Such activities may contribute to the training of the inference activities.

Based on the above, this study sets the research question: “How can we realize a learning support system that improves the ability to recognize and infer the mental states of others?” More specifically, we are interested in the following two questions: “RQ1: Does understanding the characters in dramatic scripts from multiple viewpoints contribute to the improvement of the ability to understand others?” and “RQ2: Does broadening the learner’s
thoughts based on the inferences results of others contribute to the improvement of understanding others?” In this paper, we propose a learning support system that aims to improve the ability to understand others and to cultivate an appropriate attitude toward understanding others through dramatic script reading.

2. Basic Philosophy of Fostering Ability to Understand Others

2.1 Difficulties in Understanding Others

In this study, the learning goal of the learners is to be able to appropriately recognize and infer the mental states, such as intentions and emotions, from others’ utterances and actions and to be able to think based on the other’s point of view. In order for learners to learn these skills and attitudes, the following learning difficulties need to be reduced:

- **Difficulty 1**: It is difficult to infer the mental states of others because they do not always express their emotions as they are but may intentionally control their emotions or express false emotions (Gross, 2002). It is also not easy to imagine the contextual factors that are not always expressed in the behavioral process of others because they may choose different behaviors from learners.

- **Difficulty 2**: It is not easy to consciously train oneself to acquire an attitude of trying to understand the other’s point of view without immediately judging the beliefs and actions of those who are different from oneself.

- **Difficulty 3**: It is difficult to learn desirable inferential activities because people’s mental states are complex (Berrios, 2019), and no unique inferential process can be used as a model. It is also difficult to learn from others how to perform the inference activities that take place inside a person’s mind.

2.2 Dramatic Script Reading for Developing Ability to Understand Others

In this study, we focus on dramatic scripts as a learning material for developing the ability to understand others. The act of inferring the intentions and feelings of the characters in dramatic scripts from the chain of their utterances (dialogues) and deciphering the conflicts between them is an activity to understand the intentions, feelings, and thoughts of others from the viewpoint of the characters (others) who are placed in a different background from oneself. According to Ishino, acting in theater is putting oneself in the other’s place and assuming the other’s identity, which shifts one’s point of view to the other’s (Ishino, 2015). Watanabe and Kusumi also point out that actors have the experience of carefully inferring the beliefs, desires, and motives of the characters they play to perform them naturally (Watanabe & Kusumi, 2021).

2.3 System Design for Developing the Ability to Understand Others

As an approach to RQ1, we provide an inference activity support environment that allows learners to externalize the inference results of the characters’ (others’) mental states in dramatic scripts in mind map format as an opportunity to engage in inference activities about the mental states of others. In addition, we provide “inquiries” that encourage in-depth exploration of each viewpoint. The inquiries are intended to stimulate the learners’ curiosity about the characters and further activate their critical reasoning.

The learner’s unique way of inference results will be externalized in the mind map. Therefore, as an approach to RQ2, we develop a system that captures the differences on the map between different learners and provides information that stimulates each learner’s inferential activities. This approach aims to promote learning from the inferences of others and to encourage more critical inference activities. More specifically, we aim to encourage new insights from learners by providing highlight and focus functions for each viewpoint so that learners can compare their own and others’ ideas about each viewpoint on the map. In addition, we implement an advice presentation function based on the differences between the two maps.
Advice is offered to encourage awareness of viewpoints and inquiries that the other explores in depth but has not yet explored the learner’s own and how others think about the same inquiries.

3. Learning Support System for Developing Ability to Understand Others

3.1 Environment to Support Inference Activities of Character’s Mental State

Based on the system design described in Section 2.3, we implemented a learning support system for developing the ability to understand others as a web application. Figure 1 shows the interface of the developed system. The learner can confirm the dramatic script in Figure 1(a). Here, learners grasp the flow of the entire story. After reading through the story, learners select one character they want to understand more deeply and explore the character in depth.

In Figure 1(b), viewpoints for deciphering the mental states of others (e.g., the identities of characters) are displayed in a network structure. This network is a starting point to consider the character’s mental state. The network is classified into two types: one for inferring the character’s personality (Figure 1(i)) and the other for inferring the character’s utterances and actions (Figure 1(ii)). The network for personality inference indicates stable elements related to the character’s personality development as viewpoints. The network for utterance and action inference indicates elements related to utterances or actions of the character in advance as viewpoints. By developing these networks, the learner systematically infers the character.

The system displays inquiries that delve into the understanding of the character in Figure 1(c) by clicking on a viewpoint node. For example, when the learner clicks on the viewpoint “unfulfilled desires,” adaptive inquiries (e.g., “What are the character’s unfulfilled desires?” and “Whom does the character admire?”), that delve deeper into the viewpoints are displayed. This is intended to activate critical reasoning for each point of view and to stimulate the learner’s curiosity about the character. Figure 1(d) shows a choice (e.g., ‘causality’) to consider the connection between what has been inferred under each viewpoint. This is intended to encourage reinterpreting a single interpretation from various perspectives through activities considering the connections between perspectives. In the same area, options are also displayed for a more profound investigation of the character’s thoughts (e.g., ‘reasons that led to the inference,’ ‘interpretation,’ and ‘utterances that show the interpretation’). This is intended to promote the externalization of the learner’s inference activities process.

On the interface, clicking on any node and selecting an item from these options adds a node for entering the answer below the node. The selected choice is displayed on the link, and the node can be filled in with the learner’s inferences and interpretations from the drama script. In addition, a link is created between the nodes by selecting any two nodes and

![Figure 1. Interface for Inference Activities of Character’s Mental State.](image-url)
choosing one of the options (e.g., causality). The inquiry associated with a viewpoint is displayed in a blue link, while others are in gray.

The above functions are intended to externalize the results of inferences about the mental states of characters (others) made inside the learner’s mind and the viewpoints that led to such inferences and to task the learner with critically digging into them. Currently, we set 21 viewpoints and related 42 inquiries for personality inference and 7 viewpoints and related 19 inquiries for deepening into the character from the viewpoints of the utterances and actions of characters based on the related research (Aikawa, 2015; Murayama, 1975) to personality development, human behavior, etc.

3.2 Environment to Compare Inference Activities That Differ from Learners’ Own

By selecting the other learner’s map, the learner’s and the other’s maps are displayed, as shown in Figure 2. By comparing their own and others’ maps, learners gain new viewpoints on understanding others and engage in activities to broaden their thoughts.

In the interface, when a learner clicks on a viewpoint node on his/her map, a node of the same viewpoint on another learner’s map is enlarged and highlighted in the center of the area. In the example shown in Figure 2(e), the learner selects “desire (red frame).” In addition, by selecting an utterance in the dramatic script from the area shown in Figure 1(a), the node corresponding to that line is enlarged and highlighted in the center of each map.

In addition to the visual support, the system is equipped with a function for presenting feedback (advice) focusing on the differences between the two externalized maps, as shown in Figure 2(g). The following three types of advice are generated and presented based on a template as clues to expand their thoughts by relativizing the different results of others’ reasoning activities:

- **Advice that captures differences in viewpoint**: The advice that captures the difference between the viewpoint nodes used by other learners and not by the learner. e.g., “The X (other) is deepening his/her thoughts about the character “Elizabeth’s experience” viewpoint, but you have not yet done it. Think about the viewpoint in your way”.

- **Advice that captures differences in inquiries**: The advice that captures the differences in inquiries used by other learners and not by the learner. e.g., The X (other) has been deepening his thoughts about the inquiry: “What are the character’s unfulfilled desires?” but you have not yet done it. Think about the inquiry in your way.

- **Advice that captures the same inquiries**: The advice that captures inquiries and their interpretations used by both learner and the other. e.g., The X (other) gave an interpretation as “disappointment in myself” to the inquiry: “What emotions does the character have?” Do you have any new thoughts about this?

![Figure 2. Interface for Inference Activities of Character’s Mental State.](image-url)
4. Evaluation

4.1 Experimental Setting and Procedure

An evaluation experiment was conducted to verify the usefulness of the developed system. In this evaluation, we report on the effects of the developed system on dramatic script reading toward understanding others (i.e., RQ1 and RQ2), as well as on the effects of the system on theatrical activities. Four adult members of a theater troupe and one acting instructor participated in this evaluation. The dramatic script used in the experiment was "Witch Village Forest," written by Miki Kimito. The script is 20 pages long (about 600 lines), and the screening time is about one hour.

First, the troupe members were asked to read through the script to understand the contents before using the system. No intervention, such as time control, was used here. Each member was cast with the expectation that they would play one or two characters. In order to practice map comparison activities and acting instructions after using the system, one utterance of a particular scene was designated as the target of inference activity. Then, the members were asked to conduct the inference activity (Section 3.1) of the mental state of two or three characters’ personalities (the character played by each member and the character of the dialogue partner in the dramatic script) using the system until they were satisfied with their performance. After finishing the inference activity, they were asked to conduct a comparison activity (Section 3.2) with maps created by different troupe members. The maps to be compared were the inferences of the same utterances of the same character, and they were asked to reflect on their inference results while using support functions. If any new findings emerged, they were reflected in the map.

After completing the map, the troupe members practiced acting with an acting instructor. Each member was asked to perform a part of a pre-designated scene, and the instructor was asked to guide the performance while checking the map created by the members. Finally, a questionnaire survey on a 5-point scale (Table 1) and an interview-style discussion were conducted. Here, we asked each participant to share his or her opinions on each question and asked them to share their various insights about the system.

4.2 Results

The results of the questionnaire (N=4) are shown in Table 1. Questions (A) to (D) were asked to evaluate the support function of RQ1, which aims to capture characters from multiple perspectives, and questions (E), (F), and (G) were asked to evaluate the support function of RQ2, which aims to learn from the insights of others and to broaden one’s thoughts. The results showed that all questions were generally evaluated positively, confirming that the environment for inference activities of the character’s mental state and the presentation of viewpoints and inquiries helped encourage the members to dig deeper into their characters and to perceive others from various angles.

Table 1. Questionnaire Survey

<table>
<thead>
<tr>
<th>Question</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Did &quot;viewpoints&quot; for personality and utterance/action inferences help you better understand the character from various angles?</td>
<td>4.75</td>
</tr>
<tr>
<td>(B) Did &quot;questions&quot; for each viewpoint make the activities to understand the character more active?</td>
<td>4.25</td>
</tr>
<tr>
<td>(C) Did the support function in the mind map interface help you to organize and infer the character’s mental state?</td>
<td>5.00</td>
</tr>
<tr>
<td>(D) Did a series of inferential activities help you better understand the character’s mental state?</td>
<td>4.25</td>
</tr>
<tr>
<td>(E) Did the highlighting function help you compare your thoughts with others?</td>
<td>4.25</td>
</tr>
<tr>
<td>(F) Did the advice provided by the system lead you to interpretations or insights that you had not thought of on your own?</td>
<td>4.00</td>
</tr>
<tr>
<td>(G) Did a series of activities comparing thoughts with others broaden your thoughts?</td>
<td>4.75</td>
</tr>
</tbody>
</table>
In an interview-style discussion, in response to the question to the instructor, “How has the use of the maps created by the troupe members affected your instruction?” the instructor commented that “Although I found it somewhat difficult to use in the field, it was easy to advise because I knew the actor’s thoughts in advance.” In the discussion, many members stated that using the system helped them expand their thoughts, such as “Comparing inference results of myself with the other made my thoughts more flexible” and “It became easier to revise my thoughts.” It was also commented that the system was effective in motivating the members to dig deeper into their characters, such as “I wanted to think about what others have inferred” and “Since others were able to infer more viewpoints than I could, I thought I would consider them more myself.”

In the discussion, many members stated that using the system helped them expand their thoughts, such as “Comparing inference results of myself with the other made my thoughts more flexible” and “It became easier to revise my thoughts.” It was also commented that the system was effective in motivating the members to dig deeper into their characters, such as “I wanted to think about what others have inferred.”

5. Concluded Remarks

In this study, we developed a learning support system that aims to improve the ability to understand others and to cultivate an appropriate attitude toward understanding others by using dramatic scripts as learning materials. Experimental results show that the system has the potential to facilitate inference activities that contributed to understanding others, as well as activities to infer the mental states of others from multiple viewpoints.

Future issues include improving the system’s usability and confirming its usefulness through long-term practice.

Acknowledgments

We sincerely thank the Osaka LaLaSong Musical Theater Company and its members for cooperating in this study.

References


Development of Estimation Method for Learner’s Emotional Concealment During Learning Using Biometric Information and Feedback Model

Koichi SHINOHARAa*, Keiichi MURAMATSUb & Tatsunori MATSUIc

aGraduate School of Human Sciences, Waseda University, Japan
bGlobal Education Center, Waseda University, Japan
cFaculty of Human Sciences, Waseda University, Japan
*koichis9911@toki.waseda.jp

Abstract: Although it is important to understand learners’ emotional situations in teaching and learning processes, learners sometimes do not express negative emotions intentionally. This study has termed this intentional non-expression of negative emotions “emotional concealment,” and examined its estimation from learners’ biometric information. In particular, we focused on learners’ facial features, and attempted to clarify their relationship with emotional concealment. Using these results, we attempted to quantify emotional concealment from the time-series data of biometric information and detect it as an outlier. Additionally, we attempted to develop a feedback generation system to inform teachers about emotional concealment.

Keywords: Emotional concealment, learning, estimation, feedback, micro-expression

1. Introduction

Inability to understand the content of another person’s speech is common in communication. However, continuing a conversation without resolving this situation will result in deterioration in its quality. Since the teaching-learning process can be considered communication between a teacher and learner, the same can happen during learning. In an asymmetrical relationship between the teacher and learner, the learner often does not consciously express negative feelings, such as not understanding (Shinto, 2016). The persistence of this state of mind ultimately leads to a decline in learning quality. This study terms this act of unconscious non-expression of negative emotions as “emotional concealment” (Gross, 2015). Thus, emotional concealment is defined as consciously not expressing one's negative emotions even though one is aware of them. We hypothesized that emotional concealment could be estimated using biometric information. Furthermore, we attempted to quantify emotional concealment based on biometric information and detected it as an outlier. We created a feedback model for the teacher based on the results. It is expected that the teacher will be able to understand the learner’s mental state more accurately by estimating the emotions that the learner has concealed. As a result, higher-quality learning will be realized.

2. Overview of the Study

This study aimed to develop a method for the quantification and estimation of emotional concealment using biometric information, and generate a feedback model. In Study 1, we conducted an experiment to determine whether emotional concealment occurs during
learning, and if so, what kind of biometric information it relates to. The results suggest that emotional concealment occurs during learning, and that respiration and pulse wave values are higher during emotional concealment. Additionally, facial expressions of the learners differed between emotional concealment and normal conditions. In Study 2, we conducted an experiment to investigate the differences in facial expressions during learning between emotionally concealed and normal conditions. We analyzed the intensity of facial feature points, called Action Units. Results showed that the number of Action Units around the mouth was higher during emotional concealment. In Study 3, we analyzed the time-series data of the Action Units obtained in Study 2 using an outlier detection method, assuming a relationship between the time interval showing the outlier and emotion concealment. We were able to detect anomalous values, but were unable to derive a significant relationship with the emotional concealment time interval reported in the learners’ reflection reports. The detailed methods and results of Study 4 are described below.

3. Analyzing the Feature Points of Learners’ Faces in Emotional Concealment and Normal Situations during Learning

From the discussion in Studies 1 to 3 (Shinohara, Muramatsu & Matsui, 2021), it was not possible to derive a significant relationship between the time interval showing abnormal values and emotional concealment in the Action Unit time-series data. Therefore, in Study 4, we conducted an additional experiment in which we asked the learners to watch on-demand content, and captured their facial expressions. Furthermore, we reanalyzed 67 facial landmark outputs from Open Face. The experiment was conducted with 26 Waseda University students (19 men and 7 women). As specific instructions for the experiment, learners were given the definition of emotional concealment, were told that they would be photographed during the experiment, and that they would attend the class twice, once to conceal their emotions when negative emotions occurred and the second time as usual. We obtained biometric data from learners who were watching 25-minute on-demand content on knowledge representation. We asked the learners to report whether they experienced emotional concealment during learning, and if so, at what time of day, while watching the video recordings after the experiment. Learners were asked to report their reflections on each slide of the experimental stimulus. The results showed a difference between emotional concealment and normal conditions around the mouths of all the learners (Figures 1, 2, and 3). They further suggested that the characteristic values around the mouth in the emotional concealment interval repeatedly rose and fell, as reported in all learners’ reflection reports. However, for feature points other than the lower lip, some of them, such as those that rise after falling during the emotion concealment interval, do not appear in the emotion concealment interval. Figure 4 shows a graph of the characteristic values of the upper lip. It suggests that the upper lip rises and falls not only during emotional concealment, but also during normal conditions, with no common tendency among many learners. We hypothesized a relationship between emotional concealment and micro-expressions around the mouth (Ekman & Friesen, 1987).

![Figure 1. Characteristic score graph of the lower lip of learner A.](image-url)
4. Summary and Future Tasks

This study aimed to estimate the emotional concealment state using the learner’s biometric information, and generate feedback using the results. The results of Study 4 suggest that emotional concealment is related to the learners’ facial expressions. The feature points around the mouth moved in detail when the learner was emotionally concealed. Future studies should develop a method for quantifying and estimating learners’ emotional concealment, and create a feedback model based on the developed estimation method. Using the developed feedback model, it is necessary to verify whether learning effect improves in actual learning situations.

References


Development of a Learning Support System for playing Ryuteki in Gagaku for Beginners

Yasushi UENO\textsuperscript{a*}, Masato SOGA\textsuperscript{b**}
\textsuperscript{a} Graduate School of Systems Engineering, Wakayama University, Japan
\textsuperscript{b} Faculty of Systems Engineering, Wakayama University, Japan
*\texttt{s236034@wakayama-u.ac.jp}, **\texttt{sogamasato@gmail.com}

Abstract: Ryuteki is one of the wind instruments in gagaku, a traditional music of Japan since ancient times. This instrument is difficult for beginners to acquire playing skills. The best way to learn how to play ryuteki is to learn from an expert, but the opportunity is limited due to time or cost constraints in gagaku. Therefore, in this study, we developed a system that automatically judges whether the pitch and timing of the sound are correct when a beginner plays ryuteki using sound processing technology, and presents feedback on the results.

Keywords: Learning support, Ryuteki, Gagaku, Skill, Sound processing

1. Introduction

Gagaku is a traditional Japanese music that is also called “the world’s oldest orchestra”. It was introduced from Asian countries such as China and Korea around the 5th century, and it fused with Japan’s ancient songs and dances. It was completed as a unique musical form of Japan in the Heian period, 9\textsuperscript{th} century, and it has been continuing.

One of the wind instruments used in gagaku is “Ryuteki”, which is the subject of this research. It is a transverse flute made of bamboo, and it has a similar structure to the flute in Western music (Left in Figure 1). Because it has a wide range of tones among the instruments used in gagaku, it has a role of decorating the main melody played by Hichiriki (oboe in Gagaku), as well as a role of playing “Ondo”, which is a solo performance by a Ryuteki player at the beginning of a piece.

There are various skills required to play Ryuteki, and some of them are common with the skills for playing flute, but there are many techniques and practice methods that are unique to gagaku, so it is difficult to learn gagaku performance skills even for those who are familiar with music. Especially, the differences from Western music can be seen in the notation of Ryuteki and “shoka”, which is one of the practice methods of gagaku. The notation of Ryuteki is very different from the staff notation that is often used in Western music. It has shoka written in katakana vertically in the center, and fingering written in kanji vertically on the left side (Right in Figure 1).

Shoka is a practice method of singing the melody of a piece out loud while tapping the beat with hands and knees, without using a score. By doing this, one can grasp the flow of the piece and learn how to breathe and play in that piece. Gagaku pieces are difficult to understand the flow by just looking at the score, and to be able to play them, one has to not only read the score, but also understand the flow and playing style that are not written in the score by singing shoka.

One can only practice playing with an instrument after being able to sing shoka, so there are many things that are necessary to be able to play a piece, and it can be said that it is difficult for beginners to learn how to play.

In order to learn the music of gagaku, one must grasp the flow of the music using singing and other methods, and therefore it is essential to practice under the guidance of an expert. However, currently, beginners have limited opportunities to receive guidance from instructors,
and even if they can receive guidance once, there is often a long period of time before they can receive guidance again. Even if they practice on their own during that time, it is difficult for beginners to judge whether they are playing well or not due to the characteristics of gagaku music, and therefore it is difficult for them to practice satisfactorily and efficiently improve their performance skills.

Figure 1. Ryuteki and score (Ryuteki (left) and score for Ryuteki (right))

2. Research objective

It is very difficult for beginners to practice on their own and acquire gagaku performance skills. Therefore, in this research, the learners are beginners who have acquired some knowledge about Ryuteki (a type of flute used in gagaku) and the music they want to practice, however, have difficulty performing the music with Ryuteki. The purpose of this research is to design and develop a learning support system that can help beginners judge whether they are playing correctly or not and acquire performance skills even when they practice on their own without the help of a teacher.

Specifically, we developed a system that uses audio processing technology to judge whether the pitch and timing of the performance match the musical score, and provide feedback on the results. The goal of this system is to enable beginners to judge whether they are playing “Hyojo Etenraku”, a piece that beginners often encounter first as an example of gagaku music, correctly or not on their own, and to improve their own performance by using this system.

3. Related work

As previous research on learning support systems for musical instrument performance, Takekawa et al. developed a system that uses piano fingering recognition technology to check the performer’s fingering in real time, and provides intuitive information on fingering accuracy, performance accuracy, fingering information, and keystroke information. They are necessary for performance, to enhance keystroke position and fingering proficiency in the initial stage of performance (Yoshinari Takegawa et al. 2011).

Kikukawa et al. developed a learning support system that can measure the position and angle of the bow during the performance of bowed string instruments such as erhu using a magnetic position sensor, diagnose the movement of the bow in real time, and enable beginners to acquire performance skills for bowed string instruments (Fumitaka Kikukawa, Masato Soga et al. 2014).

These studies target instruments such as piano and erhu that are played by hand or arm movements, but Ryuteki is a wind instrument that produces sound by blowing air into it internally, so it is not possible to evaluate whether it is producing correct sound or not by measuring body movements as in previous studies. Therefore, in this study, we decided to
use audio processing technology as a means of evaluating performance for wind instruments by directly evaluating the sound produced.

4. Prototype system

This system is designed for learners who have learned about the music from an instructor and want to practice on their own, so it is desirable that learners have some prior knowledge of Ryuteki performance and how to read the musical score of “Hyojo Etenraku”. We will explain how to use the system as follows.

4.1 Selecting the score to play

When a learner starts the system, a screen like Figure 2 will appear. On the left side of the screen, part of the musical score of “Hyojo Etenraku” is displayed, and by clicking on “Next part of musical score” at the bottom of the screen, the learner can select the score he/she wants to evaluate by displaying it. The frequency in the upper right shows the frequency that is currently used as the standard for judging the sound, and the learner can change the standard frequency by 1Hz in the range of 400Hz to 460Hz by clicking on the “+” and “-” buttons on the left and right. The color indicates whether the sound is not sounding (gray), in tune (green), slightly high (red), or slightly low (blue) within that scale, so the learner can check his/her own sound and tune it before actually evaluating his/her performance.

4.2 Playing the score

To practice playing a score the learner wants to play, he/she selects the score and presses “Start” to switch to the mode where he/she can play and evaluate the score. In this mode, the numbers on the left side of the screen change color from gray to white in order from top to bottom. These numbers indicate the beats of the score, and the first 3/4 are set up to prepare for playing before entering the performance, and the learner is supposed to prepare for playing during these two beats and start playing according to the score from the next beat.

The gagaku score has a length of two beats per large character of katakana singing (the part that says “Toraro…” in this score), and the learner is supposed to play along with the numbers that light up in order at regular intervals. Figure 3 shows the screen display during performance, and in this figure, it is playing the first 4th beat after starting the performance. Also, if the learner feels that his/her performance is not going well, he/she can stop playing by pressing the stop button at the bottom right of the screen, go back to the scene where he/she selects the original score, and press start again to start playing from scratch.
While the learner is playing, the system internally processes information about the sound that is currently being played and stores it in an array. This array consists of three types of arrays: an array that stores information on the scale (array 1), an array that stores the frequency deviation between the correct sound and the sound being played for that scale (array 2), and an array that stores how many seconds after pressing the start button this sound was played (array 3), and information on sound is stored in an array.

![Figure 3. Screen during performance](image)

### 4.3 Evaluation of performance

When the learner finishes playing until the last beat, the system compares the three arrays recorded during the performance with the score data and judges whether the scale matches for each beat on the score and whether there is any pitch deviation, and presents the results as feedback. Figure 4 shows an example of the screen display when giving feedback on the evaluation of performance. In this figure, it can be seen that the correct sound is produced during the first three beats, but the fourth beat is slightly high, and from there, either the sound is wrong for three beats, or the timing of switching or starting or ending the sound is off.

![Figure 4. Screen for displaying results after performance](image)

After checking the evaluation, if the learner is not satisfied with his/her performance, he/she can press the “Repeat” button to redo the performance of the same score, or if he/she feels that he/she has practiced enough, he/she can press the “Next part of musical score” button to practice the continuation of the score.

In this way, by checking from the evaluation of performance, he/she can find out which parts are easy to make mistakes or deviate in sound, and find out what he/she needs to improve in his/her performance, and by repeating practice, he/she can deepen his/her understanding of the music and practice to be able to play well. This is a system that allows him/her to practice to be able to play well.
5. Evaluation Experiment

The purpose of the experiment using this system was to investigate, using a questionnaire, whether beginners, who are the target of this study, can expect to improve their performance of “Hyojo Etenraku” by practicing with the proposed method, and whether the evaluation of Gagaku performance by the proposed method is useful for learning Gagaku. In addition, we also asked for answers on improvement points and functions that they want us to add to this system for future research.

The subjects who cooperated in the experiment were four people (two men and two women) who had experience playing Ryuteki. Of the four subjects, two were beginners, one was an expert who had performed in concerts and other events, and the remaining one was an instructor who taught Ryuteki.

In this experiment, we explained how to use the system and that it was a system built assuming that beginners would practice by themselves. We asked them to practice playing for about 10 minutes using the system. After that, we asked them to answer a questionnaire on whether they expected to improve their Ryuteki performance by using this system.

5.1 Evaluation of usefulness of the system

To evaluate the performance and functionality of the system, we conducted a questionnaire with four questions using a five-point scale. The questions and the results are shown in Table 1. The results are indicated by average points of five-point scale.

<table>
<thead>
<tr>
<th>No.</th>
<th>Question sentence</th>
<th>Average point</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>I understood which parts I played well through the system.</td>
<td>3.75</td>
</tr>
<tr>
<td>(2)</td>
<td>I understood which parts I was weak at through the system.</td>
<td>4.25</td>
</tr>
<tr>
<td>(3)</td>
<td>I felt that I could expect to improve my performance by using the system.</td>
<td>4.25</td>
</tr>
<tr>
<td>(4)</td>
<td>I felt that the method of evaluating by pitch in this study was useful for learning Gagaku.</td>
<td>4.25</td>
</tr>
</tbody>
</table>

5.2 Free comments

In parallel with the multiple-choice questions, we asked for opinions on improvement points and other things that they wanted us to incorporate in a free-writing format. A summary of the answers is as follows.

- I want them to put in a sound at the timing when the beat changes.
- I want a little more time between pressing start and starting to play.
- I felt that the judgment of pitch and tempo was too strict.
- I want them to evaluate my performance not only by the color of the numbers but also by words.
- It would be interesting to show a message by the system with a chat bot.
- I don't know if I can match with others in an ensemble, so I felt it would be nice if what I played was recorded and I could play it back and get feedback later.
- I felt that if I eliminated the Gagaku-like quality from my performance, I could get a higher score.

5.3 Discussion

5.3.1 Discussion on the usefulness of the system

From the results of the questionnaire from question number (1) to (3), it was found that by using this system, they were able to evaluate their performance, which they could not judge
when they practiced alone before, and that they expected to improve their performance. Also, since we obtained a high evaluation for question number (4), we recognize that we have obtained agreement for the method of evaluating Ryuteki performance based on pitch and timing.

5.3.2 improvements of the system

In the free comments, some users expressed their wish to have a sound cue at the timing when the beat changes, and to have a little more time between pressing the start button and starting the performance, as they felt they could not perform well in the scenario of playing. In actual gagaku ensemble, there is no conductor who keeps the beat, so players have to catch the flow of the beat by themselves, and in this system, we did not play a sound at the beat timing to create a similar situation. However, considering the opinions of the experimental results, and also considering that this system was designed with the purpose of enabling self-practice, we realized that we need to design it with more user-friendliness and playability.

6. Conclusion

In this study, we developed a prototype system of learning support for playing Ryuteki for self-learning. We evaluated the system, and we got the results that the system can support for learning Ryuteki. However, we also got opinions which tell some points to be improved. We also think the number of participants of the evaluation is only a few. We would like to increase the number of participants in future.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP 17H01996.

References

Gagaku Ryutekihu (Score for Ryuteki), Tenrikyodoyusha, 2018.


An Adaptive Learning Support System based on Ontology of Multiple Programming Languages

Lalita NA NONGKHAI\textsuperscript{a}, Jingyun WANG\textsuperscript{b} & Takahiko MENDORI\textsuperscript{c}

\textsuperscript{a} Graduate School of Engineering, Kochi University of Technology, Japan
\textsuperscript{b} Department of Computer Science, Durham University, UK
\textsuperscript{c} Graduate School of Engineering, Kochi University of Technology, Japan
* 256010u@gs.kochi-tech.ac.jp

Abstract: This research proposes to develop an adaptive ontology-based learning support system for computer programming learning. Firstly, the system adopts a previously developed ontology called CONTINUOUS, which represents programming concepts and their relation in a graph and makes use of its content to serve as hints within programming questions. Secondly, we design an adaptive strategy for recommending suitable exercises to learners, which uses CONTINUOUS as metadata of exercises and the Elo rating system to estimate learners' skills. This work aims to design a system to provide personalized exercise path for the learner and evaluate its effectiveness.

Keywords: Adaptive Learning Support, Ontology, Recommendation System, personalized learning, programming learning

1. Introduction

Computer programming utilizes coding, debugging, and evaluating of solutions (Friday et al., 2019). However, the complexity of programming problems can confuse learners regarding the necessary knowledge for the solution (Yorah et al., 2016). To improve the efficiency of learning CT (Computational Thinking), this research focuses on developing a web-based Learning Support System (LSS) based on ontology to recommend suitable programming exercises for learners. There are several educational systems leverage ontology. For instances, Ibrahim et al. (2017) utilized three ontologies to suggest suitable course for university students, and Bouihi et al. (2019) proposed an approach to apply their ontologies in recommendation system to support learners within their learning experiences by recommending appropriate materials.

In our previous work (Lalita et al., 2022), we developed an ontology named CONTINUOUS, encompassing common concepts from Python, Java, and C#, describing programming concepts, sub-concepts, and concept relation. This research adapts CONTINUOUS to visualize programming concepts as a graph to help learners understand the related concepts in a specific programming language. Furthermore, the design principle of CONTINUOUS determines knowledge points (KP) as minimal syntax of a specific programming language and enables KP to serve as a hint of each programming question to guide learners in solving a specific problem.

Moreover, this research also focused on designing a personalized recommendation approach to enhance learner programming skills. Previously, Zheng et al. (2022) have devised a knowledge structure tree for C programming, leveraging it as meta-data in their recommendation system. They used the matching degree calculation to match score between the student's cognitive level and exercise difficulty to facilitate personalized exercise recommendations. Similarly, Michlík et al. (2010) have employed linear extrapolation to estimate potential knowledge levels in exercise recommendations, with the goal of aiding learners in achieving favorable exam results within a limited time. To enhance the assessment of learner's understanding level, we adapt the Elo Rating System (ERS) (Elo, 1978), which is
a method used to assess and measure the relative proficiency of the participants in
competitive games and has also been utilized in education research to estimate learner’s skill
and item difficulty (Radek, 2016). Additionally, Klinkenberg et al. (2011) demonstrated the
efficacy of ERS in assessing player skill and item ratings within the Math Garden, a
mathematics game for primary students.

In summary, the contribution of this research is the adoption of CONTINUOUS in an LSS
and the design of an adaptive strategy for exercise recommendations to explore whether this
LSS can effectively enhance learners’ programming skills.

2. Development of an LSS Based on Ontology in Computer Programming

2.1 The adoption of CONTINUOUS with the system

Our system utilizes CONTINUOUS to visualize programming concepts as a graph to guide
learners with a clear understanding of the relation between concepts. For instance, as shown
in Figure 1, if learners decide to learn the “data_structure” concept in Python, they have the
option to practice with “List” or “Dictionary.” Opting to study the “List” concept necessitates
understanding of the associated “list_method”, which serves as a function for manipulating
elements within the “List” concept.

In addition, the contents in CONTINUOUS are also used as a hint in programming
questions, depending on the code answer in each question. Currently, our system contains
questions from a question bank, which has 68 questions in three levels of difficulty (easy,
standard, and difficult) involving concepts in CONTINUOUS. For example, when learners are
tasked with verifying the equivalence of input numbers, which involves the concepts of
“conditionals” for decision-making and “standard_input” for keyboard input, the system will
assist learners in grasping essential concepts (“conditionals” and “standard_input”) for
addressing the question.

2.2 The design of adaptive strategy to recommend suitable programming exercises

To recommend suitable programming exercises for learners, the system will update a
learner’s skill based on the formulation of ERS presented by Radek (2016) in an educational
setting as follows:

\[ \theta_s := \theta_s + K \cdot (\text{correct}_{st} - P(\text{correct}_{st} = 1)) \]  (1)
From formulation 1, we denote $\theta_s$ as the learner’s skill of a learner $s$, the value of $K$ as a learning rate, which is a weight value in the formulation, $\text{correct}_{si}$ as the correctness of code answer of a learner $s$ on an exercise $i$ and $P(\text{correct}_{si} = 1)$ as the probability of a correct answers as shown in formulation 2:

$$P(\text{correct}_{si} = 1) = \frac{1}{(1 + e^{-(\theta_s - d)})} \quad (2)$$

The initial value of $\theta_s$ is set to 0. After code submission, the system evaluates it with the testcases. If all testcases are passed, $\text{correct}_{si}$ is set to 1; otherwise, it is set to 0. Learner’s skill is estimated in the range of 0 to 1 and compared within three difficulty levels of exercises. If the result is nearly 0 indicates a slight increase of maintain the same, while the result close to 1 which means learner’s skill increase, leading to the system providing the upper level. If learners reach the highest level within the current concept, the system will show the practice process of one of its related concepts to learners.

3. Conclusion and Future Work

This research focuses on developing an LSS, which adopts CONTINUOUS to visualize programming concepts as a graph, enabling learners to understand the relation between concepts within a specific language. Additionally, the system makes CONTINUOUS serve as exercise metadata and uses the programming concepts of CONTINUOUS as hints to support learners during problem-solving. Furthermore, this research designs an adaptive strategy for exercise recommendations, utilizing the ERS formulation to estimate and update learners’ skills. In the future, we plan to conduct several experiments to evaluate the system and answer the research questions: whether our system which provides personalized exercise recommendation based on individual understanding levels, can effectively enhance learners’ programming skills. Learning log data will also be collected and analyzed.

References


Examination of the robot's role as a helper in learning situations

Makoto SHIRAISHI\textsuperscript{a}, Tatsunori MATSUI\textsuperscript{b}
\textsuperscript{a}Graduate School of Human Sciences, Waseda University
\textsuperscript{b}Faculty of Human Sciences, Waseda University

Abstract: In this study, we aim to realize Learning by Teaching, in which a person learns by teaching a robot in a learning situation between the robot and the person. In this paper, we report an experiment to investigate the impression of a helper on a recipient in a general learning situation between humans. We focused on the robot's behavior that makes the learner want to teach the robot, concentrating on its weaknesses. We then conducted an additional experiment in which we changed the assisting person from a human to a robot to realize the robot's behavior model. In the future, we will compare these two experiments to examine the behavioral model.

Keywords: Weak robots, supporting behavior, HRI

1. Research Background and Purpose

In recent years, an increasing number of robots have been introduced in the field of education, and various effects have been reported. Han (2008) suggested that robot-assisted English language learning is more effective than textbooks or computer-based learning in terms of children's concentration, interest, and learning (Han, 2008). In addition, a study conducted by Matsuzoe (2013) suggested that a robot's presentation of correct answers and teaching was effective in learning English words (Matsuzoe, 2013). The role of the robot is to be taught, which affects not only the effectiveness of the learner's learning, but also the learner's motivation to learn. Therefore, this study aimed to realize learning-by-teaching, in which the learner learns by teaching a robot. To achieve this goal, this study examines the role of robots in teaching and focuses on weak robots. A weak robot performs actions after eliciting assistance from others (Okada, 2016). It would be useful to design such learning by giving the robot a role, such as weakness, to encourage the learner to teach, which would result in the completion of the learning process. However, previous studies have been vague in their definitions of weakness. In this study, weakness was defined as the desire to help. Therefore, we conducted two surveys to verify what kind of weaknesses can trigger learning-related assistance in learning situations. By comparing these two experiments, the weaknesses of the robot in learning situations were clarified.

2. Experiment 1

As the first step of Experiment 1, we surveyed aid workers' impressions of their aid recipients in person-to-person learning situations.

2.1 Experiment Participants and Tasks

An online questionnaire survey was conducted during the experiment. The participants were asked to respond to episodes in general learning situations regarding their experiences with assistance, reasons for assistance, and impressions of the assisted person. There were 100
participants in the experiment (61 males and 39 females), with a mean age of 40.62 years (standard deviation, 9.15).

The purpose of the experiment was to use a questionnaire to ask participants to answer questions about episodes of assistance in each common learning situation. In the assistance episode, which was an experimental stimulus, the learning situation applied to small, everyday acts of kindness (compassion). A brief survey was conducted on "Please tell us about your compassion during education and learning." This survey and the Compassion Scale by Kikuchi (1988) were used as references to create the assistance episodes. Participants were asked to respond on a 5-point scale (ranging from “never did” to “did more”) to record their experience assisting others, based on the Compassion Scale. Participants were also asked to provide a free description of the reasons for their assistance. As for the impressions of the aided persons, the participants were asked to respond to 48 items such as “bright” to “dark” and “strong to weak,” from the adjective pairs in the validated scale for measuring personality perception, as well as “smart” to “dumb” and “approachable to inapproachable,” for a total of 50 items (Hori & Yoshida, 2001).

2.2 Result

In this experiment, data from 97 of the 100 subjects were used for the analysis. The data were analyzed using exploratory factor analysis for each of the 15 assistive episodes. Initially, for each aid episode, we used the Kaiser-Meyer-Olkin measure of sample validity. The results of the measurement showed a result of 0.87 or higher, for all assistance episodes, indicating the possibility of a potential factor. Based on the above results, factor analysis was conducted by unifying the maximum likelihood method for factor extraction. The results identified 10 factors for the 15 assistance episodes. These factors were activity, sociability, diligence, reliability, politeness, affinity, agreeableness, reactivity, robustness, and impatience. Overall, each factor had a high value for the factor loadings of the items included in the factor, high explanatory power for "introverted," "static," etc., and low as an explanation for the factor (Table1).

<table>
<thead>
<tr>
<th>Item</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic - Static</td>
<td>0.799</td>
<td>0.168</td>
<td>0.185</td>
</tr>
<tr>
<td>Lively - Lonely</td>
<td>0.788</td>
<td>0.103</td>
<td>0.264</td>
</tr>
<tr>
<td>Energetic - Tired</td>
<td>0.782</td>
<td>0.227</td>
<td>0.227</td>
</tr>
<tr>
<td>Smart - Stupid</td>
<td>0.331</td>
<td>0.828</td>
<td>0.146</td>
</tr>
<tr>
<td>Decent - Sloppy</td>
<td>0.255</td>
<td>0.827</td>
<td>0.268</td>
</tr>
<tr>
<td>Calm - Restless</td>
<td>0.069</td>
<td>0.819</td>
<td>0.239</td>
</tr>
<tr>
<td>Like - Dislike</td>
<td>0.224</td>
<td>0.373</td>
<td>0.762</td>
</tr>
<tr>
<td>Kind - Strict</td>
<td>0.021</td>
<td>0.208</td>
<td>0.746</td>
</tr>
<tr>
<td>Honest - Obstinate</td>
<td>-0.005</td>
<td>0.313</td>
<td>0.711</td>
</tr>
</tbody>
</table>

2.3 Discussion

After conducting the experiment, 10 factors were identified for all 15 episodes. Each factor tended to be low, suggesting that the helper felt some kind of weakness toward the assisting person. It is possible that each factor was characterized by a mature impression of the activity factor and a shy impression of the sociability factor. In many episodes, there were also descriptions of sympathy, such as “because I felt poor,” and of concern, such as “because they looked like they were in trouble.” This is considered the same tendency of emotional quality as in previous studies on helping behavior (Nakamura & Takagi, 1987). Weaknesses have been suggested to be involved in learning situations. Additionally, the free descriptions indicated that most participants assumed a friend-like person. This study also attempted to use robots as learning partners. This was done to capture similar trends. Therefore, an additional experiment was conducted by replacing the assistant with a robot.
3. Experiment 2

To conduct Experiment 2, I revised the presentation of the assumed situation and assistance episodes and the aid episodes to fit the robot scenario. The episodes utilized in Experiment 1 were abstracted and thoroughly discussed in the laboratory to create episodes adapted to circumstances applicable to the robot.

3.1 Experiment Participants and Tasks

Similar to Experiment 1, an online survey was conducted for Experiment 2. Participants were asked to respond to a general learning episode regarding the frequency of assistance, reasons for assistance, and impressions of the recipient of assistance. There were 104 participants (59 males and 45 females), with a mean age of 39.98 years (standard deviation, 8.79).

In Experiment 2, the same questionnaire as in Experiment 1 was used, and the participants were asked to respond to questions about episodes of assistance with the robot in each general learning situation. The questionnaire was designed in the same manner except that the questions were changed to the frequency of assistance.

3.2 Result

As in Experiment 1, an exploratory factor analysis was conducted on the 15 assistance episodes, and a likewise procedure for factor analysis was followed. Nine factors were identified, namely activity, sociability, intelligence, reliability, sincerity, favorability, robustness, clarity, and warmth. Overall, each factor tended to have a high factor loading value for the items included in the factor and, as in Experiment 1, a low explanatory value. The structure of each factor was similar to that in Experiment 1, suggesting that the respondents felt the same sense of weakness as in Experiment 1.

3.3 Discussion

Similar to Experiment 1, the results of Experiment 2 suggest a sense of weakness in the learning situations. Among these, sociability, intelligence, reliability, and favorability factors were identified in many episodes. Many factors with similar structures were identified, but more personality-related factors were identified in Experiment 2. This can help realize the robot's behavior. However, in this experiment, we presented only images of the robot and were unable to verify its behavior. In the future, the impressions confirmed in the experiment will be reflected in a behavioral model. This information has been included in the poster presentations.

References

A Bio-Inspired Method for Personalized Learning Path Recommendation Problem

Yaqian ZHENG\textsuperscript{a, b}, Deliang WANG\textsuperscript{c}, Yaping XU\textsuperscript{a, b}, Ziqi MAO\textsuperscript{a, b}, Yaqi ZHAO\textsuperscript{a, b} & Yanyan LI\textsuperscript{a, b}*

\textsuperscript{a}School of Educational Technology, Beijing Normal University, China
\textsuperscript{b}Research Center for Knowledge Engineering, Beijing Normal University, China
\textsuperscript{c}Faculty of Education, The University of Hong Kong, Hong Kong
*liyy@bnu.edu.cn

Abstract: The recommendation of personalized learning paths is recognized as one of the most challenging aspects in the field of e-learning. In the existing literature, numerous approaches have been proposed to identify appropriate learning paths for e-learners, taking into consideration multiple perspectives. However, the current state of research lacks a unified framework that effectively integrates the most vital parameters associated with the learner, learning object (LO), and domain knowledge to generate optimal learning paths. To address this challenge, a novel bio-inspired approach is proposed for solving the personalized learning path problem. In this method, we initially incorporate the learner, LO and domain knowledge models into a unified mathematical model. Then an enhanced ant colony optimization algorithm is utilized to determine the optimal personalized learning paths for learners. To investigate the effectiveness of the proposed method, we performed several computational experiments based on six simulation datasets. The results indicate that the proposed method surpasses other competing methods in terms of performance and robustness, showcasing its superior effectiveness.

Keywords: E-learning, learning path recommendation, unified model, ant colony optimization algorithm

1. Introduction

E-learning has gained significant prominence in the education sector, providing remarkable opportunities for learning (Al-Fraihat et al., 2020). However, e-learners, particularly those with limited experience in a specific domain, may face challenges such as information overload and learning disorientation when confronted with a wide range of resources (Meng et al., 2021). In order to tackle this issue, several methods have been proposed for the automated generation of personalized learning paths for e-learners. These methods aim to optimize learning paths from various perspectives, including domain knowledge structure and learning preferences, among others. Nevertheless, there is currently no unified framework that effectively incorporates the most significant parameters linked to the learner, learning object (LO), and domain knowledge in order to create optimal learning paths. To this end, a novel bio-inspired approach is proposed to discover personalized learning paths for e-learners, in which the important parameters associated with learner, LO and domain knowledge are formulated as a unified mathematical model and an enhance improved ant colony optimization (EACO) algorithm is utilized to identify the optimal personalized learning path.

2. Problem Formulation

The personalized learning path recommendation problem (PLPRP) is represented by a three-tuple consisting of a learner model, LO model, and knowledge graph. Figure 1 presents a description of the problem formulation.
Here, the PLPRP is considered as a constrained optimization problem. The unified mathematical model is expressed as follows:

$$\text{Minimize } F = w_{\text{match}}F_{\text{match}} + w_{\text{continuity}}F_{\text{continuity}}$$ (1)

Subject to

$$\text{KP}_{\text{set}}(G') \subseteq \text{KP}_{\text{set}}(P)$$ (2)

$$(\text{KP}_i, \text{KP}_j) \in E, \forall \text{KP}_i, \text{KP}_j \in \text{KP}_{\text{set}}(P), 1 \leq i < j \leq \text{KP}_{\text{number}}(P)$$ (3)

where $F_{\text{match}}$ assesses the matching degree between the learner and LOs based on Learner-LO matching rules and $F_{\text{continuity}}$ evaluates the path's continuity based on the dependency relations between KPs. Constraint (2) is employed to ensure the path's completeness, while constraint (3) is utilized to guarantee its legality.

3. A Bio-Inspired Method for Solving the PLPRP

To solve the model of PLPR, a bio-inspired algorithm, i.e., an enhanced ant colony optimization algorithm (EACO) is proposed.

- Problem representation. The problem space is transformed into a KP space and LO space. Once an ant selects a node, the KP and LO space undergo dynamic updates. When the KP space becomes empty, a feasible learning path solution is constructed.
- Heuristic information. The matching degree between the learner and LO is leveraged to design the new heuristic information.
- Pheromone initialization. Following prior work (Yu et al., 2018), the value of pheromone is initialized by a greedy nearest-neighbor algorithm.
- Dynamic candidate set strategy. To generate feasible solutions of learning path, a dynamic candidate set strategy is presented, which constructs the candidate set for each element based on the dependency relations.
- Selection of the initial node. The heuristic information and the roulette selection rule are utilized to determine the initial positions of the ants.
- Selection of the next node. The selection of the next node is determined using the pseudorandom-proportional rule (Ma et al., 2021).
- Pheromone Update. After an ant finishes constructing a solution, the pheromone levels on the edges visited by the ant in the obtained solution are locally updated. Once all the ants have completed their solutions, the pheromone trails are evaporated, and only the best-so-far ant is permitted to release pheromone on all the edges along its traveled path.

4. Computational Experiments

We generate multiple simulation datasets to perform computational experiments and assess the effectiveness of our proposed algorithm. The datasets are defined based on the number of KPs and LOs. The specific settings are as follows: (10, 30), (15, 45), (25, 75), (40, 120), (60, 180), and (90, 270). For the purpose of comparison, three widely used algorithms, namely random algorithm (RA), genetic algorithm (GA), and standard ant colony optimization
algorithm (SACO), are employed. Considering the inherent stochastic nature of these algorithms, each of them is independently executed 20 times for every dataset.

Table 1 provides a summary of the computational results, presenting the average objective function value over 20 runs (labeled as \( \text{Obj} \)) and the standard deviation of the objective function values across different trial runs (labeled as \( \text{Std} \)). The values of \( \text{Obj} \) obtained by RA are always the highest, so its performance is the poorest. This indicates that learners encounter challenges when attempting to independently explore suitable learning paths. GA, SACO, and EACO obtain better solutions than RA, demonstrating that employing heuristic algorithms can effectively guide the search process toward more promising areas. In all scenarios, EACO demonstrates the highest level of performance, and its superiority becomes increasingly prominent as the problem dimension expands. In comparison to other algorithms, RA exhibits significantly larger values of \( \text{Std} \), which imply its limited stability. On the other hand, IACO consistently achieves the smallest \( \text{Std} \) values across all scenarios, highlighting its remarkable robustness.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RA ( \text{Obj} )</th>
<th>RA ( \text{Std} )</th>
<th>GA ( \text{Obj} )</th>
<th>GA ( \text{Std} )</th>
<th>SACO ( \text{Obj} )</th>
<th>SACO ( \text{Std} )</th>
<th>EACO ( \text{Obj} )</th>
<th>EACO ( \text{Std} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2530</td>
<td>0.1745</td>
<td>0.5339</td>
<td>0.0514</td>
<td>0.3430</td>
<td>0.0188</td>
<td>0.3250</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>1.1626</td>
<td>0.1589</td>
<td>0.5141</td>
<td>0.0585</td>
<td>0.3106</td>
<td>0.0199</td>
<td>0.2250</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>1.2734</td>
<td>0.1728</td>
<td>0.9649</td>
<td>0.0426</td>
<td>0.7035</td>
<td>0.0231</td>
<td>0.2170</td>
<td>0.0024</td>
</tr>
<tr>
<td>4</td>
<td>1.3364</td>
<td>0.1497</td>
<td>0.9888</td>
<td>0.0224</td>
<td>0.6915</td>
<td>0.0210</td>
<td>0.1966</td>
<td>0.0065</td>
</tr>
<tr>
<td>5</td>
<td>1.6351</td>
<td>0.1262</td>
<td>1.3335</td>
<td>0.0474</td>
<td>1.0830</td>
<td>0.0199</td>
<td>0.1308</td>
<td>0.0022</td>
</tr>
<tr>
<td>6</td>
<td>1.7216</td>
<td>0.1375</td>
<td>1.1698</td>
<td>0.0159</td>
<td>1.1670</td>
<td>0.0342</td>
<td>0.2112</td>
<td>0.0059</td>
</tr>
</tbody>
</table>

5. Conclusions

In this work, we propose a bio-inspired algorithm to address the personalized learning path recommendation problem (PLPRP). The experimental results indicate that compared with other competitors, the proposed algorithm achieves better performance and robustness when dealing with the PLPRP across varying sizes and levels of complexity.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant No: 62277006), the Beijing Natural Science Foundation (Grant No: 9222019), the Science and Technology Research Youth Fund Project of Higher Education Institutions of Hebei Province (Grant No: QN2018261) and the Student Research Fund of the Faculty of Education at Beijing Normal University (Grant No: 2212216).

References

Developing Effective Educational Chatbots with ChatGPT prompts: Insights from Preliminary Tests in a Case Study on Social Media Literacy

Cansu KOYUTURK**, Mona YAVARI†, Emily THEOPHILOUb, Sathya BURSICa, Gregor DONABAUERa,c, Alessia TELARId, Alessia TESTAa, Raffaele BOIANOd, Alessandro GABBIAKIT, Davinia HERNANDEZ-LEOb*, Martin RUSKOV†, Dimitri OGNIBENEa,f***

aUniversità of Milano Bicocca, Milan, Italy  
bUniversitat Pompeu Fabra, Barcelona, Spain  
cUniversity of Regensburg, Regensburg, Germany  
dPolitecnico di Milano, Milan, Italy  
eUniversità of Milano, Milan, Italy  
fUniversity of Essex, Colchester, United Kingdom

*davinia.hernandez-leo@upf.edu, **martin.ruskov@unimi.it, ***Lead PI dimitri.ognibene@unimib.it  
†Same contribution

Abstract: Educational chatbots come with a promise of interactive and personalized learning experiences, yet their development has been limited by the restricted free interaction capabilities of available platforms and the difficulty of encoding knowledge in a suitable format. Recent advances in language learning models with zero-shot learning capabilities, such as ChatGPT, suggest a new possibility for developing educational chatbots using a prompt-based approach. We present a case study with a simple system that enables mixed-turn interactions and discuss the insights and preliminary guidelines obtained from initial tests. We examine ChatGPT’s ability to pursue natural educational conversations, adapt the educational activity to users’ characteristics, such as culture, age, and level of education, and its ability to use diverse educational strategies and conversational styles. Although the results are encouraging, challenges are posed by the highly structured form of responses by ChatGPT, as well as their variability, which can lead to an unexpected switch of the chatbot’s role from a teacher to a therapist. We provide some initial guidelines to address these issues and to facilitate the development of effective educational chatbots.

Keywords: ChatGPT, Educational chatbots, Educational prompts

1. Introduction

Educational chatbots have gained increasing attention in recent years for their potential to provide interactive and personalized educational activities with reduced teacher intervention. However, their development has been mostly focused on chatbot-driven conversation flow due to the complexity of mixed-initiative and user-driven approaches, the restricted interaction capabilities of available platforms, the lack of adequate training sets (Pérez et al., 2020), and the substantial efforts required to encode the necessary knowledge in a suitable format, notwithstanding the wide availability of topic-specific information.

Recent advances in Large Language Model (LLM) systems with their zero-shot learning capabilities (Wei et al., 2021), also named prompt-based approaches (Xu et al., 2023), to perform diverse tasks in different domains without training or fine-tuning suggest a new way to develop educational chatbots. Chatbots based on Reinforcement Learning from Human Feedback (RLHF), such as ChatGPT, have been extended to achieve more reliable and safe interactions (Ouyang et al., 2022).
While showing big promises, LLM-based chatbots and their potential are not yet fully understood (Mahowald et al., 2023; Borji, 2023) and present different sources of vulnerability (Xu et al., 2023). The skills necessary for natural educational conversations, such as playing a specific role, maintaining a conversation style (Lu et al., 2023), and assessing students’ conditions (Ullman, 2023), are still challenging even for state-of-the-art systems.

We propose an investigation framework to study ChatGPT’s ability to produce mixed-turn educational conversations where the students can steer the conversation toward desired topics as well as cover multiple related learning objectives, tailor educational activities to user characteristics such as culture and age and use different educational strategies and conversational styles. Building on Jeon and Lee (2023), in this paper, we provide suggestions for possible teacher orchestration and interrogation strategies and study whether they lead to successful interlocution and content provision. We explore different prompting strategies to test LLM capabilities in educational conversations with a particular focus on reliability, interactivity, and coverage. We report results from tests performed by the authors.

2. Methodology

To frame interactive educational chats with ChatGPT, we focused on the "space of learning" (Marton and Tsui, 2004). We explored the educational exchange during lessons, especially the challenge of imperfect communication between teachers and students. Adding the chatbot as a teacher mediator (Fig. 1), we interpreted the learning interaction using an adaptation of Laurillard’s Conversational Framework (2009), considering the distinctions between conceptions and expressions for each role (teacher, student, and chatbot-mediator).

![Figure 1. The educational framework: The teacher's intention to convey learning objectives is expressed through generated prompts in the technical framework. The chatbot's stochastic model determines the communication directed to the student. As the student interprets the exchanged messages, their understanding takes shape. Consequently, the actual experience of the messages differs from the teacher's intention, the chatbot's representation, and the student's comprehension.](image)

To establish a learning space and facilitate a directed educational exchange we have formulated six blocks of educational objectives to be integrated into the prompt. Each block targets a critical area of social media literacy (Ognibene et al., 2023; Theophilou et al., 2023). The assessment of the interaction quality between ChatGPT and users was evaluated based on four dimensions: correctness, consistency, conversation fluency, and semantic coverage.

3. Results and Conclusions

We report: I. the insights that emerged during our search for prompts that resulted in effective educational conversations, and II. a classification of the issues we encountered. For details see Koyuturk et al. (2023). Our study showed that ChatGPT with the correct prompt could play the role of the teacher in short interactive educational conversations covering multiple learning objectives related to social media literacy. We found that it is useful and functional to explain in the prompt every foreseeable aspect of the interaction one by one. We aimed to clearly define the structure of interaction and the role of the chatbot. By testing with eight personas with different cultural backgrounds and ages, we verified that ChatGPT could also adapt to students’ characteristics. However, relatively small variations of the prompt lead to a range of issues (see table 1).
ChatGPT appears to be fine-tuned for question-answering, resulting in responses that may not align with the desired style or exhibit essential conversational elements. Initial prompt attempts resulted in unnatural response formats. Despite instructions to adopt a conversational style and avoid bullet points or essay formats, the chatbot frequently deviated. Turn-taking and questions were either lacking or excessive and the ability to cover multiple learning objectives and related topics was limited. Conversations often continued only when users asked relevant questions about previously unaddressed topics.

Interesting erratic and undesirable behaviors were observed. Requests to make the educational conversation interactive sometimes resulted in the failure to effectively maintain the teacher role. One of the most surprising and frequent behaviors was the tendency to autonomously cover both sides of the educational conversation. Our work shows that studying and extending the LLM’s capabilities to consistently play roles is an essential line of research.

Acknowledgements
Spanish AEI (PID2020-112584RB-C33, MDM-2015-0502), Volkswagen S. (ref. 95 566 and 9B145) and Catalan SGR (00930). DHL (Serra Húnter) acknowledges the support of ICREA Academia.

References
Tackling Unserious Raters in Peer Evaluation: Behavior Analysis and Early Detection with Learner Model

Changhao LIANG*a, Izumi HORIKOSHIb, Rwitajit MAJUMDARb & Hiroaki OGATAb

*a Graduate School of Informatics, Kyoto University, Japan
b Academic Center for Computing and Media Studies, Kyoto University, Japan

* liang.changhao.84c@st.kyoto-u.ac.jp

Abstract: Peer evaluation of individual or group work is often adopted in team-based learning design. However, some raters may not take the evaluation process seriously and exhibit behaviors such as using the same score, rushing through evaluations, or not evaluating during the presentation. This study investigates the issue of unserious peer evaluation in group presentations, focusing on their behavior patterns. Using evaluation behavior analysis indicators, we identified unserious raters who exhibited low reliability in the peer evaluation process. Further, we conducted a preliminary analysis to detect unserious raters based on learner model data available before the peer evaluation process. This information can assist teachers in providing personalized prompts and interventions prior to the peer evaluation process, thus enhancing the evaluation quality of these students with timely prompts to them.

Keywords: Peer evaluation, Evaluation behavior analysis (EBA), Team-based learning (TBA), Data-driven study, Learner model

1. Introduction

Evaluation is an essential aspect in collaborative learning, but teachers may struggle to properly evaluate each student (Amarasinghe et al., 2021). Peer evaluation offers formative feedback that encourages reflection and overcomes the limitations of traditional evaluation (Ohland et al., 2012). It has become widely adopted in online settings where student-centered learning is prevalent and can enhance both learning and interpersonal skills (Kasch et al., 2021). However, some raters may not take the evaluation process seriously, as Horikoshi and Tamura (2021) discovered. Such evaluations may involve using the same score, rushing through evaluations, or not evaluating during the presentation. These low-quality ratings can make peer evaluation results less reliable and lower the learning outcome.

Nevertheless, the reliability of unserious peer raters can be improved by proper interventions (Van Zundert et al., 2010). Current studies have made attempts to calibrate scores based on student engagement and previous performance (Piech et al., 2013), or train evaluation skills during the peer evaluation process (Gorham et al., 2023). These approaches can be too late to nudge timely interventions to the ongoing evaluation activity. Fewer work addresses predicting problem raters early before the assessment to facilitate possible interventions to improve their evaluation behaviors.

This study investigates the issue of unserious peer evaluation in group presentations, focusing on their behavior patterns. Using behavioral indicators, we identified unserious raters who exhibited low reliability in the peer evaluation process. Further, we conducted a preliminary analysis to examine how the learner model data from their learning logs and their prior peer evaluation behaviors can be used for early detection. This information can assist teachers in providing personalized prompts and interventions prior to the peer evaluation process, thus enhancing the evaluation quality of these students in a timely manner.
2. Research Background

2.1 Peer evaluation in Team-Based Learning (TBL)

In peer evaluation, students provide ratings and feedback on each other's work, which is formative and can promote their performance in subsequent tasks (Ohland et al., 2012; Gorham et al., 2023). Research has shown that peer evaluation encourages students to think deeply and critically about their own work and contributes to the development of "internal feedback" skills, where learners reflect on and regulate their own learning processes (Nicol et al., 2014; To & Panadero, 2019).

Team-Based Learning (TBL) is an educational strategy that involves multiple rounds of group work with peer evaluation. It was first introduced in medical education (Michaelsen et al., 2002). During each round of TBL, students start by exploring the learning topic individually before working in teams to complete tasks (Parmelee et al., 2012). Group presentations and peer evaluation conclude each round, where students assess the products or outcomes of their peers' learning experiences and reflect themselves as a formative process (Topping, 1998). Moreover, under the data-driven environment, previous rounds' learning log data enables targeted interventions by teachers (Johnson, 2017).

2.2 Evaluation Behavior Analysis (EBA)

The process of peer evaluation generates behavior indicators that record key information, such as the identity of the evaluator, the timing of the evaluation, the items assessed, and the corresponding scores (Horikoshi & Tamura, 2021). The behavior indicators stem from "paradata" in the web survey research field, which refers to the log data generated during the evaluation process and is related to the quality of survey responses (Couper & Kreuter, 2013). For instance, shorter response times are associated with a "lack of motivation to answer accurately caused by continuous survey" (Yan & Tourangeau, 2008), and individuals who answer quickly as "speeders" can lead to poor responses (Zhang & Conrad, 2014).

The web survey research and peer evaluation research share the goal of measuring inappropriate behaviors in digital evaluation platforms. Therefore, to effectively analyze and visualize the quality of peer evaluation based on behaviors, the Evaluation Behavior Analysis (EBA) method has been developed. It involves extracting data from peer evaluations and utilizing it to gain insights into students' evaluation behaviors. Using the EBA method, instructors can identify patterns and trends in the evaluation behavior of students. Horikoshi et al. (2022) have defined feature variables that capture the key aspects of evaluation behavior, which are presented in Table 1.

<table>
<thead>
<tr>
<th>Behavior Indicator</th>
<th>Definition</th>
<th>Proposed constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Time (ET)</td>
<td>Time span from clicking the first evaluation item to the last item.</td>
<td>Speed: how much time the rater spent on the evaluation</td>
</tr>
<tr>
<td>Mean of the Timestamp (tM)</td>
<td>Average elapsed time since the start of the presentation.</td>
<td>Timeliness: whether the rater evaluated immediately after the presentation</td>
</tr>
<tr>
<td>SD of the Timestamp (tSD)</td>
<td>Standard deviation of the timestamps for all evaluations.</td>
<td>Coherence: whether the rater evaluated evenly throughout or within a short time</td>
</tr>
<tr>
<td>Click Count (CC)</td>
<td>Total number of times the evaluation items were clicked.</td>
<td>Certainty: how many changes the rater made</td>
</tr>
<tr>
<td>Mean of the Score (sM)</td>
<td>Average score for all the evaluation items scored by the reviewer</td>
<td>Leniency: rater tendency to assign higher or lower scores</td>
</tr>
</tbody>
</table>
Compared to the conventional perspective of peer evaluation quality, which primarily emphasizes scores and compliance with others (Cho & Schunn, 2007; Fukazawa, 2010), the EBA indicators focus on the process of the peer evaluation. These evaluation behavior indicators go beyond mere consistency and provide insights into various aspects of peer evaluation performance (see Proposed construct in Table 1). Such insights can inform instructional design and enable targeted feedback. By identifying strengths and weaknesses in specific behavior indicators, EBA allows for the recognition of areas that need intervention, thus promoting the development of peer evaluation skills in line with the goal of formative assessment and learning enhancement.

### 2.3 Data-driven peer evaluation with learner model

Peer evaluation systems offer learners a scaffold to evaluate their group members and receive real-time feedback with reduced bias, enhanced individualism and privacy protection (Ismail et al., 2016; Cleynen et al., 2020). In online evaluation systems, both the evaluation outputs and the evaluation processes of raters can be traced, providing valuable data for learning analytics applications as part of the learner model attribute. The concept of the "learner model" encompasses domain-specific and domain-independent information, quantified as learning evidence that varies according to the learning context (Boticki et al., 2019). These indicators can derive from learning behaviors recorded on learning management platforms (LMS) such as e-book reading logs, academic scores, previous experiences in group work, and other relevant data. In the context of TBL, the learner model can be dynamic and continuously updated with the accumulation of data from each round. To support this process, Group Learning Orchestration Based on Evidence (GLOBE) (Liang et al., 2021) was proposed as an infrastructure that provides data-driven support for group work based on learner model data. Peer evaluation plays a significant role in GLOBE, serving as a module for collecting peer ratings and feedback (Liang et al., 2022a), while also contributing to the modeling of effective group work and task experiences (Janssen & Kirschner, 2020). By synchronizing the evaluation data with other collaboration attributes from the prior phase, the learner model can be utilized for subsequent rounds of TBL.

The data-driven perspective has been adopted to assess the quality of peer evaluation in individual tasks. For instance, Piech et al. (2013) developed tuned models of rating reliability based on students' previous performance in individual design assignments. Besides, there are studies focusing on written reviews for writing artifacts. Cho & Schunn (2007) considered consistency with others to model reviewers' capabilities, while Patchan et al. (2016) extracted features from review texts, such as sentimental tendencies and comment types, using semantic analysis to build a regression model. Regarding peer evaluation in group work, Liang et al. (2022a) demonstrated that the accumulated learner model, incorporating data on group work and task experiences, can estimate the consistency of peer evaluation using GLOBE. However, for iterative TBL with multiple rounds of group work, the detection of evaluation behaviors on rating scores has yet to be extensively investigated.

### 3. Method

In this study, we conduct two analyses to answer the research questions. First, we examined the behavior patterns of unserious raters based on evaluation behavior indicators using clustering analysis. Then, to investigate the potential of using the learner model data from learning logs for early detection, a preliminary analysis of classification was conducted. The research questions are as follows.

- **RQ1**: What are the peer evaluation behavior patterns?
3.1 Participant and context

The data of this study comes from a course of a Japanese university with a 4-week experiment. The course is for students beyond sophomore in computer science, with 35 students enrolled this year. The experiment employed an adapted TBL and jigsaw design, which is shown in Figure 1.

In the first week of the experiment, a lecture on a new topic was delivered and BookRoll, an e-book reading tool that allows instructors to upload learning materials before each class and enables students to engage in various activities during their reading (Ogata et al., 2015), was introduced. Out-of-class activities included reviewing lectures on BookRoll, participating in forum discussions, and completing assignments to summarize them. Starting from the second week, in-class activities began with group sharing of the previous week’s assignments. Each student presented the outcome from their forum discussion group in a jigsaw group. Both the forum discussion groups and jigsaw groups were created by the group formation system of GLOBE (Liang et al., 2021) based on each student’s reading engagement in BookRoll. In each jigsaw group, the audience provided peer ratings on the individual presentation through the peer evaluation system. The jigsaw group then became the forum discussion group for the following week. Following this, a lecture on the topic of the second week was delivered. This workflow was repeated twice in the first three weeks, and as an assignment in the third week, students created a presentation to the whole class, summarizing what they had learned so far and presented it in the final week’s class. The behavior pattern analysis in this study is based on the peer evaluation of this final presentation.

3.2 Data collection and preprocessing

In order to evaluate the final group presentation, students were instructed to assign a score on a 5-star scale to each group in the peer evaluation system (Liang et al., 2022a). The rubric was displayed at the top of the rating section in the peer evaluation system for reference (see Figure 2). The system also recorded a log of the timestamp and rating score each time a rating button was clicked by a student. To ensure privacy, the identity of each student was anonymized from the log. Using the clicking logs, six evaluation behavior indicators introduced in Table 1 were calculated. These indicators are used for visualization and clustering.

To detect unserious behavior prior to the peer evaluation, data from the learner model was collected. In this study, the following learner model data was available before the final peer evaluation of presentations:

- **Reading engagement (RE)**, which includes reading time, operation times, completion rate, and the number of red markers, yellow markers, and memos on the e-book platform BookRoll (Ogata et al., 2015). A comprehensive coefficient was calculated by averaging the percentage rank of the aforementioned indicators to represent reading engagement.
Forum engagement (FE), which consolidates the number of forum posts and characters in the out-class forum discussion. The percentage rank of the former indicators was consolidated to represent the forum engagement.

Prior evaluation behavior indicators, which refer to indicators collected during the peer evaluation of individual presentations in jigsaw groups in the second (round 1) and third week (round 2). The six indicators introduced in Table 1 were collected for the first two rounds as the input indicators for classification.

As some of the prior evaluation behavior indicators were found to be highly correlated and estimating the same facet, as also mentioned in Horikoshi et al. (2022), we performed dimension reduction through factor analysis. Based on the factor analysis, we combined ET and tSD as “time feature (TF)” (explaining 99.01% of variance for round 1 and 98.63% for round 2), and sM and sSD as “scoring feature (SF)” (explaining 84.91% of variance for round 1 and 91.97% for round 2). Additionally, since the extent of polarization in tM was deemed important in the pattern, we derived a new indicator (tDEV) from tM, which represents the z-score of tM and describes the deviation of rating time from the mean. CC is treated as an independent indicator due to its low correlation with other features. We used eight prior behavior indicators (four for each round) for the classification modeling.

3.3 Data analysis

To answer RQ1, we performed a clustering analysis to differentiate unserious raters from the participants. This analysis entailed clustering the students according to their evaluation behavior indicators, which were obtained from the final round of peer evaluation (for group presentation). We utilized the K-means method to obtain two distinct clusters, with the highest silhouette score. Subsequently, we examined the behavior patterns of the students by analyzing the distributions of each evaluation behavior indicator within each cluster.

For RQ2, we approached it as a binary classification problem to determine if the rater is unserious in evaluating the final group presentation. To accomplish this, we tested five commonly used machine learning classification models for numerical data and evaluated their
performance using the Area Under Curve (AUC) (Fawcett, 2006), with values ranging from 0 to 1. Furthermore, we conducted a feature ablation analysis (Gabrilovich & Markovitch, 2004) based on the information gain (IG) of ten input indicators as discussed in the previous section, to figure out the predictive indicators for the classification.

4. Result

4.1 Behavior patterns clustering

Figure 3 illustrates the distribution of EBA indicators for each cluster. It is evident that students in cluster C1 possess longer ET, more CC, and give a wider range of scores with lower sM and higher sSD. Although tM does not show a significant difference between the two clusters, the dispersion differs. Raters in C1 participated in peer evaluations during the presentation, and their distribution of timestamps appears to be more normalized. On the other hand, students in cluster C2 have shorter ET, fewer CC, polarized tM, and smaller tSD. Regarding scores, they tend to provide full marks, indicated by high sM and minimal sSD.

Table 2. Statistical analysis of EBA indicators for clustering

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Cluster</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>ET</td>
<td>C1</td>
<td>21</td>
<td>41.716</td>
<td>11.259</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>14</td>
<td>9.658</td>
<td>10.748</td>
<td></td>
</tr>
<tr>
<td>tM</td>
<td>C1</td>
<td>21</td>
<td>58.870</td>
<td>8.416</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>14</td>
<td>50.775</td>
<td>19.494</td>
<td></td>
</tr>
<tr>
<td>tSD</td>
<td>C1</td>
<td>21</td>
<td>17.137</td>
<td>5.670</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>13</td>
<td>5.173</td>
<td>5.401</td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>C1</td>
<td>21</td>
<td>6.762</td>
<td>2.791</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>14</td>
<td>4.071</td>
<td>1.492</td>
<td></td>
</tr>
<tr>
<td>sM</td>
<td>C1</td>
<td>21</td>
<td>3.941</td>
<td>0.508</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>14</td>
<td>4.777</td>
<td>0.378</td>
<td></td>
</tr>
<tr>
<td>sSD</td>
<td>C1</td>
<td>21</td>
<td>0.746</td>
<td>0.370</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>13</td>
<td>0.113</td>
<td>0.218</td>
<td></td>
</tr>
</tbody>
</table>

***p < .001.
Early detection of unserious raters

Figure 4 presents a performance comparison of various classification methods when using the top N input indicators ranked by IG, and Table 2 listed these indicators in the order of their IG in the classification modeling. Our analysis suggests that neural network and logistic regression models outperform other methods when utilizing the top five to six input indicators with high information gains. The AUC scores were 0.738 for the 5-feature condition (Neural Network) and 0.731 for the 6-feature condition (Logistic Regression).

As for predictive indicators, we observed that the deviation rating timestamp for round 2, indicating a straightlining pattern, had the highest IG. Additionally, SF for both rounds exhibited high information gains. Interestingly, all four prior behavior indicators for round 2 ranked in the top six indicators of the classification model. We also observed a significant difference between the two groups in SF for round 1 and TF for round 2. Meanwhile, the reading behaviors of the two groups that occurred before the assessment started. The RE feature also provided valuable information for distinguishing between different classes in a classification, underscoring the importance of integrating learning model data in predictive modeling. Conversely, the tDEV, CC, and TF of round 1 had low IG, which could be attributed to the unfamiliarity with the system in the first round as students needed time to get accustomed to it.

Table 2. Input indicators for the classification modelling ranked by information gain

<table>
<thead>
<tr>
<th>Rank</th>
<th>Indicator</th>
<th>Information Gain (IG)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tDEV-2</td>
<td>0.226</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SF-2</td>
<td>0.211</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>RE</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SF-1</td>
<td>0.178</td>
<td>3.251**</td>
</tr>
<tr>
<td>5</td>
<td>CC-2</td>
<td>0.154</td>
<td>1.976</td>
</tr>
<tr>
<td>6</td>
<td>TF-2</td>
<td>0.077</td>
<td>2.498*</td>
</tr>
<tr>
<td>7</td>
<td>tDEV-1</td>
<td>0.071</td>
<td>0.974</td>
</tr>
<tr>
<td>8</td>
<td>FE</td>
<td>0.055</td>
<td>1.264</td>
</tr>
<tr>
<td>9</td>
<td>CC-1</td>
<td>0.049</td>
<td>0.773</td>
</tr>
<tr>
<td>10</td>
<td>TF-1</td>
<td>0.031</td>
<td>0.397</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01.
5. Discussion

The findings of this study emphasize the significance of integrating learning model data in peer evaluation of TBL. By using EBA indicators, we can analyze the time and scoring features of peer evaluation as the presentation progresses. These indicators can reveal behavior patterns suggested by Horikoshi & Tamura (2021) such as modifying the evaluation, spending time on the evaluation, or evaluating all evaluation items earlier or many evaluation items later. The clustering corroborates these patterns and identifies characteristics of unserious raters.

To provide a clearer understanding of the behavior pattern, Figure 5 displays plots of the evaluation behavior of typical raters in the two clusters, indicating the timestamps, scores, and rating intervals. The x-axis represents the elapsed time from the start of the first group presentation, and the y-axis denotes the candidate number of peer ratings. It can be observed that typical students in C1 tend to rate each candidate group across the group presentation sessions with even intervals. Moreover, they use different scores with noticeable variations. In contrast, typical students in C2 exhibit a straightlining and speedy pattern (Zhang & Conrad, 2014; Kim et al., 2019), completing their rating very quickly either at the beginning or the end of the session. In summary, C1 raters spend more time evaluating their peers, give a diverse range of scores with less agreement among themselves, and exhibit a more even distribution of timestamps when giving their evaluations. C2 raters, on the other hand, spend less time evaluating their peers, give higher scores with less variance, and show a polarized distribution of timestamps for their evaluations. These differences suggest that C1 raters demonstrate more thoughtful and critical evaluations, while C2 raters appear to be more lenient and less engaged in the evaluation process.

Moreover, this study presents the potential of using learner model data collected from all phases of TBL in previous rounds to early predict unserious raters. Our analysis shows that scoring features in each round of TBL play a significant role in the detection model. Time features, which describe the time distribution and frequency of the ratings, can also be predictive when TBL is conducted over multiple rounds and raters become familiar with the system. Furthermore, the engagement of students in individual reading activities can serve as a predictor of unserious raters, while their forum engagement appears to be less relevant. This discrepancy may be attributed to forum posts being compulsory and formatted as part of the course grade, resulting in minimal variation among learners. In summary, the prediction model is expected to empower instructors to provide remedial instructions or give automatic nudges to these at-risk students, improving the reliability of peer evaluation as a formative assessment in TBL. These prompts can be delivered through group awareness tools (Strauß & Rummel, 2021) and email interventions (Damgaard & Nielsen, 2018).

The study also contributes to TBL design by introducing the potential of data-driven scaffolding with multiple rounds of group learning, where learning log data from previous rounds can be utilized for various learning analytics purposes. One example of data-driven...
support is provided for the detection of unserious raters in peer evaluation, which is intended to improve the quality of peer assessment. Furthermore, continuous data support has broader applications in group learning. The learning logs of previous activities can provide data for creating groups (Liang et al., 2022b) and calibrating peer rating scores (Piech et al., 2013). The accumulated data can also be useful for data visualization platforms for reflecting on teaching interventions (Kuromiya et al., 2020).

However, there are several limitations to this study. The sample size of learners observed was relatively small, which might limit the generalizability of the findings. Moreover, it should be noticed that the current predictive model's AUC did not achieve a high level, and the model needs to be validated using a different student population. Besides behavior indicators, we plan to incorporate the consistency of the ratings, including the agreement with instructor-assigned grades and average student-assigned grades (Fukazawa, 2010), into the prior evaluation behavior indicators. Further, considering more predictors in the model, such as learning outcomes, collaborative skills, and personality variables (Piech et al., 2013; Sánchez et al., 2021), could also enhance its effectiveness. Qualitative observations and self-reports can offer valuable insights into the reasons behind unserious patterns, and exploring how the presented EBA estimated from logs connects to the observations is another promising topic. Lastly, since this research only involved one trial of a group presentation, conducting additional studies with more rounds of TBL and peer-evaluated group presentations is anticipated to address remaining issues and enhance the robustness of the findings.

6. Conclusion

In conclusion, this study discusses the issue of unserious raters in peer evaluation of group learning. We propose a method to describe unserious peer raters by detecting trends based on the clustering of EBA indicators. The results reveal typical behavior patterns of unserious raters: straightlining, speeding, and giving all full marks. Next, a preliminary evaluation is conducted for classifiers that can identify groups of unserious raters. The results revealed typical time and scoring features associated with these raters, as well as predictive indicators for early detection. Overall, these findings have implications for improving the effectiveness and reliability of peer evaluation in group learning contexts. Further investigation is required to explore the actual quality of ratings and validate the classification model.

Acknowledgements

This research was supported by the following grants: JSPS KAKENHI 20K20131, 20H01722, 22H03902, NEDO JPNP20006, JPNP18013, and JST JPMJSP2110.

References


Future City: A Simulation for Making SDGs Action Decisions

Pin-Chen CHEN*, Yu-Hao LU, Ju-Ling SHIH
Graduate Institute of Network Learning Technology, National Central University, Taiwan
*daisy12296021@gmail.com

Abstract: The Sustainable Development Goals (SDGs) deliver a concept aimed at addressing social, economic, and environmental challenges that require collaborative efforts from everyone. The actions often involve conflicts of interest between businesses, governments, social organizations, and citizens. This study develops an inquiry-based learning system <Future City> aiming to illustrate SDGs issues. Students play different roles and simulate future social scenarios based on their positions and action decisions. This assists students in understanding the concepts of sustainability.

Keywords: SDGs, Simulation, Action Decisions

1. Introduction

To enhance students' awareness of sustainable development, the Taiwanese government has integrated the United Nations' proposed SDGs into education which involves not only videos and reading materials but also activities that encourage student participation and reflection. The effectiveness of these efforts goes beyond mere promotion; they are primarily focused on education. As a result, this system builds upon the <City Auncel> developed by Lu, Shih, & Hong (2022), and further develops an issue simulation decision system designed to engage students in various societal roles. Students work within the interests of their respective roles while also considering the common welfare. This entails finding a balance between ecological environments and socio-economic development, achieved by resolving conflicts of interest through effective negotiations. Consequently, this study delves into the impact of SDGs simulations on learners' understanding of the SDGs and their behavioral decisions.

2. Literature Review

2.1 SDGs-related simulation

The United Nations the SDGs to mitigate climate change and promote sustainable development on a global scale (Rosati & Faria, 2019). However, establishing partnerships among businesses, governments, and civil society can be challenging if they have competing goals (Brinkerhoff & Brinkerhoff, 2011).

Computer simulation is a tool used to approximate real-world results and aid in future planning. For example, Colliste (2017) utilized the iSDG dynamic model to simulate the annual effects of solar energy investment and policy interventions on the SDGs of Good Health and Well-Being, Quality Education, and Affordable and Clean Energy. The study investigated the causal relationship between electricity acquisition, education, and life expectancy, and the results showed that improved electricity supply was associated with increased educational attainment and life expectancy. Another simulation system, iplan (Ruis et al., 2023), is based on urban development actions and uses a complex computational model to simulate land use scenarios, helping players explore and solve land use problems and cultivate students to make meaningful actions to address social and environmental issues.

This study focuses the development of a simulation system based on the survival crisis of the Formosan Leopard Cat, which is related to sustainable cities and communities, climate action, and life on land. In the simulation, students assume different roles, including...
government officials, conservationists, developers, hunters, and farmers, to explore issues related to urban development and ecological conservation.

2.2 Action Decision

In the simulation, students had to make developmental decisions. During the process, the current social environmental situation is taken into consideration. Students collect information, evaluate the uncertainty of action consequences, and analyze the pros and cons of various actions before making the best next possible decision. Different decision models emerge depending on the tasks at hand, and the decision maker possesses analytical characteristics considering the variables that appear in the field. Students evaluate the probability of future states and target schemes with the focus on the decision event (Klein & Calderwood, 1991). In this study, each player’s decision-making process will be analyzed in regard of the inquiry-learning and issue-discussion.

3. System Architecture

The simulation system supports the inquiry process in three stages: Inquiry, Action, and Negotiation, that are cycled for three rounds (Figure 1).

![Figure 1. Simulation process flow chart](image)

The first stage, Inquiry, requires students to select a region for exploration, identify problems relating to SDGs issues, and provide explanations. The second stage, Action, students choose a development action, and the system simulates the environmental changes caused by the action. The third stage, Negotiation, requires students to resolve conflicts of interests arise from the developmental actions done by various roles in future cities. All behaviors in the simulation are synchronized, presented, and recorded in the database.

4. Simulation Design

Roles in the simulation has respective independent values and missions that the collective action would affect the common geographical values (Table 1).

Table 1. Independent and common missions

<table>
<thead>
<tr>
<th>Missions</th>
<th>Independent</th>
<th>common</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government</td>
<td>Public opinion&gt;60, Environment &gt; 60, Public opinion &gt; 60, Public category &gt;=1.</td>
<td>Development&gt;60</td>
</tr>
<tr>
<td>Conservationists</td>
<td>Environmental protection&gt;=2, Environment &gt; 60</td>
<td>Environment &gt;60.</td>
</tr>
<tr>
<td>Developers</td>
<td>Buildings &gt;=2, Money &gt; 1000.</td>
<td></td>
</tr>
<tr>
<td>Hunters</td>
<td>Production&gt;50, Money &gt; 300.</td>
<td></td>
</tr>
<tr>
<td>Farmers</td>
<td>Production&gt;50, Money &gt; 300.</td>
<td></td>
</tr>
</tbody>
</table>

There are 20 actions players can choose from, including public, conservation, construction, hunting, and agriculture (Figure 2). All actions cause corresponding numerical value changes. For instance, building an industrial area upstream of a river or in a general...
plain area make different impacts on water pollution. Thus, it is important that the player investigates the relevant data of an area before taking action. Action labels are added to the map with the identification of respective roles (Figure 3). Players can inspect the impact of their actions on the environment and adjust their future action decisions.

![Figure 2. Action option diagram.](image1)

![Figure 3. Future City Simulation.](image2)

5. Methods

This study plans to use pre-test and post-test as well as questionnaires to understand students’ attitudes, awareness, and knowledge of SDGs. The collected system data will be statistically analyzed to understand whether their understanding of the concept and perceptions of sustainable development goals has improved.

6. Expected Results

Utilizing an inquiry-based learning approach, this research empowers students to delve into real-world data, ensuring that they possess a substantial background knowledge before making behavioral decisions. The study aims to incorporate SDGs issues through role-play simulations, investigating students’ understanding and attitudes towards addressing SDGs problems. It also examines their ability to access relevant information and their decision-making process in response to simulated future social and environmental changes. The negotiation processes they engage in while resolving conflicts of interest among different groups will be analyzed. By scrutinizing the disparities between actual data and simulated data resulting from behavioral decisions, users can reflect on the environmental impact they have generated, which aids in their comprehension of SDGs issues. The simulation system is anticipated to promote the understanding of SDGs issues through an entertaining approach and encourage players to allocate more attention to the topic of sustainable development.

References


Multimodal Assessment of an Ultra-Brief Practice of Progressive Muscular Relaxation Adapted for the Classroom

Marc BEARDSLEY*, Batuhan SAYISab & Marta PORTERO-TRESSERRAc

aDepartment of Information and Communication Technologies, Universitat Pompeu Fabra, Spain
bDepartment of Computer Science and Technology, University of Cambridge, United Kingdom
cDepartment of Psychobiology and Methodology of Health Sciences, Institute of Neuroscience, Universitat Autònoma de Barcelona, Spain

*marc.beardsley@upf.edu

Abstract: Progressive muscular relaxation (PMR) is a technique often used in clinical settings for reducing physical symptoms of anxiety and improving well-being. This paper presents a comparison of psychological and physiological effects of an ultra-brief, video-guided PMR activity (uPMR) that has been adapted for classroom settings and a time-matched control activity (CA); and student opinions of uPMR to gauge its acceptability for use in formal educational settings. Forty undergraduate students (18-25 years) from a Spanish university participated in the study in which measures of state anxiety (State-Trait Anxiety Inventory, STAI), affect (Affective Slider, AS), and parasympathetic nervous system (PNS) activity indexed via the heart-rate variability parameter root mean square of successive differences (RMSSD) were conducted. Results show that RMSSD was significantly higher during uPMR than when measured before. The same effects were not found with CA. Additionally, participants rated the design of uPMR more favorably and acceptable for use in formal educational settings than CA. Significant changes in affect and state anxiety; and sustained physiological effects were not detected, suggesting that the brevity of the activity may dampen its effects. The findings related to the physiological effects and student opinions support the notion that uPMR possesses promising characteristics for use in situated teaching and practice of self-regulatory techniques aimed at reducing physiological arousal on an acute basis. Hence, it can be considered an appropriate activity to integrate into technologies that support teachers in creating opportunities for self-regulation in the classroom and guiding students through evidence-based self-regulation practices.

Keywords: Self-regulation, Physiology, Stress, Affect, Progressive muscular relaxation

1. Introduction

“Globally, the mental health of university students is an important public health issue” (Bantjes et al., 2022; p. 809) and cost-effective, sustainable, and scalable interventions are needed. A systematic review of review-level evidence on supporting university student mental health, reports on a wide range of interventions. Some of the interventions with evidence of positive effects on student mental health include skill-training programs with supervised practice, stress management, relaxation training, mindfulness-based interventions, interventions delivered via technology, and settings-based interventions which aim to enhance academic teaching and learning (Worsley et al., 2022).

Stress is a driver of mental health challenges. University students experience high levels of stress which affects their health, academic performance, and longer-term employability (Pascoe et al., 2020). There is a need to educate students about the effects of stress and to introduce them to efficient coping strategies both to support student academic
performance and well-being (Vogel & Schwabe, 2016). Relatedly, emotional self-regulation, defined as the ability to influence our own emotions or those of others, is an important self-management skill. It involves recognizing one’s emotional state, selecting, and implementing a regulatory strategy to modify it, and monitoring for success in achieving the regulatory goal (McRae & Gross, 2020). Self-regulation strategies such as mindful breathing, progressive muscular relaxation (PMR), and reappraisal can decrease symptoms of anxiety and down-regulate negative emotions (Worsley et al., 2022; Scult et al., 2017).

Building upon the idea that formal educational settings can provide students with safe and situated opportunities for learning and practicing self-regulation, this work contributes to efforts that use technology to increase educator access to resources and techniques that do not require specialized training to teach (i.e., they can be delivered by a non-expert), do not take a long period of time to demonstrate (i.e., they can be integrated into a regular class), and can help educators better provide groups of students opportunities to improve abilities to self-regulate. Along these lines, a digital technology for teaching and learning was developed to support teachers in creating opportunities for self-regulation in the classroom and guiding students through evidence-based self-regulation practices. The application (ClassMood App) involves students reporting their current emotional state and based on an aggregate state of the class, the teacher selects and runs an activity to alter the emotional state of the class (Beardsley et al., 2019). ClassMood App aims to help students become aware of their own emotions and experience actions they can take to regulate them; and also, to become aware of the emotional states of their peers and the overall class – as the classroom emotional climate can affect class performance (Brackett et al., 2011). To increase accessibility, ultra-brief (2 to 3-min), video-guided activities have been integrated into the application. Obradović et al. (2021) found evidence that an ultra-brief (1-min), video-guided deep breathing intervention effectively reduced children’s physiological arousal and was an accessible tool that could be used in home, school, or online settings. Activities available in ClassMood App include evidence-based techniques such as mindfulness and breathwork. However, the psychological and physiological effects of ultra-brief variations of these evidence-based techniques is unclear. This work aims to investigate the effects of an ultra-brief, video-guided practice of progressive muscular relaxation (uPMR).

PMR is a stress management technique often used in clinical settings that involves the repeated contracting and relaxing of different muscle groups. It has also been used for reducing physical symptoms of anxiety in the general population (Klanin-Yobas et al., 2015), in children (Vagnoli et al., 2019), and among university students (Gallego-Gómez et al., 2019). PMR works by promoting a deep state of relaxation, activating the parasympathetic system (Sakakibara et al., 1994), decreasing cortisol levels (Gallego-Gómez et al., 2019), and decreasing brain activity (Kobayashi & Koitabashi, 2016). A typical PMR intervention consists of multiple 20-min sessions across 7–10 days (Wilczyńska et al., 2019). Several studies of PMR have been run involving university students. Schienle and Unger (2021) studied a 14-day PMR course for university students, in which participants were to practice PMR each day. Participants reported lower levels of arousal, more positive valence, higher relaxation levels after their daily PMR practice and lower stress levels at the end of the course. Pawlow & Jones (2005) ran a single 20 to 25-min PMR practice and compared the effects to a no-treatment control. The authors found that PMR practice lowered levels of heart rate, state anxiety, perceived stress, and salivary cortisol. The authors (Dolbier & Rush, 2012) of a study on a single, 25-min, lying down PMR practice with high stress university students, concluded that PMR can decrease cognitive anxiety, increase mental and physical relaxation, and enhance restorative physiological functioning. The latter is represented by an increase in parasympathetic nervous system (PNS) activity which, in turn, is indicated by changes in heart-rate variability (HRV). Dolbier and Rush (2012, p. 51) write that HRV refers to changes in the time intervals between consecutive heartbeats with greater fluctuation indicating PNS activity, whereas lower fluctuation indicates sympathetic nervous system activity. The latter studies suggest a single, abbreviated practice of PMR can have positive psychological and physiological effects on university students.

Aligning with settings-based interventions that offer skill-training related to stress management with the support of technology, this present study investigates whether uPMR is
acceptable for the situated teaching and practice of self-regulation in formal education. To do so, the psychological and physiological effects of uPMR on student anxiety, affect, and parasympathetic nervous system (PNS) activity indexed via the HRV parameter root mean square of successive differences (RMSSD) are assessed in comparison to a time-matched control activity (CA). RMSSD is thought to reflect cardiac vagal activity which depicts PNS contribution to cardiac regulation (Laborde et al., 2017). RMSSD is widely used in emotion regulation studies (Balzarotti et al., 2017) Moreover, student opinions of uPMR are also presented to assess student acceptance of the activity. Hence, there are three hypotheses being investigated in this paper. H1: uPMR will lead to more beneficial physiological effects than a time-matched control activity (CA) as evidenced by a greater increase in HRV (RMSSD) from a comparison of baseline and activity performance periods (Pawlow & Jones, 2005; Nickel et al., 2005; Dolbier & Rush, 2012). The difference in HRV reflects higher post-pre PNS activity. H2: uPMR will lead to more beneficial psychological effects than (CA) as evidenced by lower pre-post ratings of arousal and higher valence (Schienle & Unger, 2021), and lower pre-post ratings of state anxiety (Pawlow & Jones, 2005; Dolbier & Rush, 2012). H3: University students will rate uPMR as more acceptable for use in formal educational settings than a time-matched control activity that involves watching a video documentary.

2. Methodology

2.1 Study Design

The protocol is a conceptual replication of the work by Laborde et al. (2022) that compared the effects of variations of a paced-breathing activity using physiological (HRV, RMSSD) and psychological measures (anxiety, arousal, and valence). Periods of rest, reactivity, and recovery following the 3Rs of HRV (Laborde et al., 2017) underlie the structure of the protocol. Rest is the baseline; reactivity is the change between baseline and the event (i.e., self-regulatory activity); and recovery is the change between the event and a post-event measure. Modifications to the replicated protocol include the use of a Vanilla Baseline (VB) to replace the resting forced relaxation (i.e., as the recovery task). For VB, participants perform a task requiring sustained attention but minimal cognitive load (Jennings et al., 1992) to better reflect a students’ classroom experience. For this study, participants were asked to count the number of times different colors appeared on a screen. The color changed every 6 s and the task lasted 2 minutes. Further, videos were used to guide participants through the self-regulation activities. This follows the approach of Obradović et al. (2021) to mimic how educators would use materials that have been produced in a scalable manner (i.e., showing students a video adapted for classroom use).

2.2 Participants

There were 40 participants (22 female, 18 male) from 18 to 25 years old ($M = 19.60 \pm 1.64$). Psychophysiological data from 7 participants were excluded due to technical issues related to electrode movement on signal integrity (6) and misunderstanding of self-report measures (1). The final sample for the psychophysiological data was 33 participants (18 female, 15 male) from 18 to 25 years old ($M = 19.48 \pm 1.67$). All participants were compensated 15 euros for their participation and gave written informed consent.

2.3 Materials and Measures

In this within-subject design, participants joined two 1-hour laboratory sessions. The sessions followed the protocol shown in Figure 1 in which two self-regulation activities (uPMR or alternatives) and one control activity were performed. Each activity lasted 2 min and there was a 6-min break between the performance of activities. To control for carryover effects between
activities, the sessions took place on separate days. The current study presents a comparison of one of the self-regulation activities (uPMR) and a neutral control activity (CA).

![Figure 1. The Experimental Protocol](image)

2.3.1 Ultra-Brief Progressive Muscular Relaxation Activity (uPMR)

A publicly available video (https://youtu.be/CATd9fqb-Mg) that was created in a multidisciplinary European project was used to guide participants through the 2-min uPMR activity. In the video, an animated character leads viewers through a cycle of tensing and releasing muscles starting from the feet, thighs, and hands to the shoulders and face. The video was edited to omit the introduction to PMR and only included a 2-min clip to guide the performance of PMR.

2.3.2 Control Activity (CA)

A 2-min clip from a documentary (“Abenteuer Forschung” [Research Adventures]) on outer space was shown to participants as the control activity. This video was found to have emotionally neutral ratings and was used in a previous study on the autonomic effects of breathing exercises (You et al., 2022).

2.4 Procedure

During the welcome phase, participants were introduced to the sequence of activities (see Fig. 1). Trial tasks and activities were run on PsychoPy (v2021.2.3) and shown to participants on a monitor. As participants completed a survey collecting demographic information and data related to possible confounds (Laborde et al., 2017), sensors were placed on them (sensor attachment phase). During the familiarization phase, the self-report scales (STAI-S, AS), resting task, VB task, and self-regulation activities were introduced, and participants were able to ask questions to clarify tasks. For the resting task, participants were instructed to look at a black screen with a fixed white cross at its center (Parchment et al., 2016). Progression of the trial for the self-report questionnaires was controlled by the participant whereas the resting task, VB task, and activities advanced automatically based on pre-set times. The trial ended with a debriefing phase in which participants completed a survey of their views of the activity, the sensors were removed, and debriefing occurred. Ethical approval was received for the adapted protocol.

2.4.1 Dependent Variables

Current state of anxiety was measured using a 6-item, short form of the State-Trait Anxiety Inventory (STAI-S) (Mateau & Bekker, 1992). Student affect was measured using the Affective Slider (AS) which consists of a continuous scale with icon images on either end to measure levels of emotional arousal (“How active do you feel?”) and emotional valence (“What is your mood?”) (Betella & Verschure, 2016). HRV was computed from the raw electrocardiogram (ECG) signal acquired from a wearable for capturing physiological data during the performance of full-body movements (Sayis et al., 2022). The wearable records the
ECG signal via 5 electrodes (2 on the chest, 2 on the shoulders, and 1 placed on the neck). Data was acquired at a sampling rate of 500 Hz. An event-based synchronization approach was followed to synchronize data sources (Bannach et al., 2009).

2.4.2 Post-Trial Survey

Participants completed a Google form survey with closed-ended and open-ended questions at the end of each laboratory session. This paper reports an analysis of the closed question, a 7-point Likert scale (1: strongly disagree, 7: strongly agree) that gathered student opinions on 11 items for each activity (see Table 2). Also, an analysis of an open-ended question related to the PMR activity that asked participants to specify if there was anything preventing them from relaxing during the activity is presented.

2.5 Data Analysis

The use of physiological (HRV) and self-report measures (AS, STAI-S) follows the approach of (Laborde et al., 2022). ECG signal data were imported into Kubios (version 2.2) and analyzed manually for artifacts. Features were extracted in time and frequency domains and nonlinear indices included in Kubios. Checks for normality and outliers were done. The RMSSD data were non-normally distributed, so a log-transformation was applied (InRMSSD) consistent with HRV research (Laborde et al., 2017). For AS, valence and arousal data were mostly normally distributed. However, 1 outlier was removed from valence ratings. STAI-S data were non-normally distributed, so log-transformation (InSTAI-S) was applied. A series of repeated-measures ANOVA with Greenhouse–Geisser correction were conducted with conditions (uPMR vs. CA) and time (PRE, DURING, POST) set as independent variables. HRV data for the full 2-min duration of each period (PRE, DURING, POST) was used. Valence, arousal, level of anxiety were self-report dependent variables, and InRMSSD was an HRV-dependent variable. Adjustments for multiple comparisons were done with Bonferroni correction. Descriptive statistics were also used for the survey data. A paired-samples t-test was used to determine whether there was a statistically significant mean difference in views toward activities. Statistical analyses were computed using SPSS (v29).

3. Results and Discussion

Descriptive statistics are presented in Fig. 2 for InRMSSD, Table 1 for AS (arousal and valence) and InSTAI-S. In all figures, * represents $p \leq 0.05$, ** represents $p \leq 0.01$, *** represents $p \leq 0.001$.

3.1 H1: Physiological Effects of uPMR in Comparison to CA

**HRV (InRMSSD):** A repeated-measures ANOVA confirmed that there was no overall significant difference between the two conditions, $F(1, 32) = 3.47, p = 0.071$. However, there was a significant interaction effect for condition $\times$ time, $F(1.85, 59.48) = 29.08, p < 0.001$. Pairwise comparisons revealed higher levels of InRMSSD during the uPMR condition compared to the levels registered during the CA ($p < 0.001$). Moreover, no statistically significant differences between conditions in InRMSSD levels during PRE ($p = 0.785$) and POST ($p = 0.229$) were found. Interestingly, pairwise comparisons in uPMR condition showed significantly higher levels in InRMSSD during the session compared to baseline levels measured prior ($p = 0.044$). Nevertheless, no significant differences in InRMSSD levels comparing PRE- and POST-measures in uPMR were observed ($p = 0.420$). These results suggest that uPMR increases InRMSSD levels compared to the CA while the activity is being performed, however this increase does not appear to show long-lasting effects.

Our findings differ from those of Nickel et al. (2005) as the authors found that RMSSD values were higher post-activity following a 25-min PMR practice in a clinical population.
However, the pattern of results found in our study are similar to the ones observed in You et al. (2022) where the increase in InRMSSD was also found during a brief (17-min), video-guided paced-breathing activity, but not maintained after the intervention. Other studies did not show changes in physiological measures such as heart rate, HF-HRV or respiration after a brief PMR training (Smith & Norman, 2017). Our results confirm a significant increase in PNS activity associated with the relaxation response during uPMR performance (Hoffman et al., 1982). Nevertheless, cardiac vagal activity recovered to baseline after practice. A possible reason could be that participants move immediately after the activity to answer the self-report surveys, causing PNS activity to decrease. A more likely explanation is that the intervention time was too brief to have a longer lasting impact on PNS activity. Moreover, muscle relaxation, in general, might have greater muscular effects and smaller autonomic effects than other relaxation methods such as finger temperature biofeedback and/or autogenic training (Lehrer et al., 1994).

![Figure 2. Descriptive Statistics (InRMSSD)](image)

### 3.2 H2: Psychological Effects of uPMR in Comparison to CA

**Affect (AS):** Regarding emotional arousal, there was no significant main effect of condition, $F(1, 32) = 1.036, p = 0.316$; no interaction effect for condition × time, $F(1.62, 51.83) = 1.306, p = 0.275$. While uPMR DURING and POST values were lower than PRE levels, the differences were not significant, whereas CA POST values were significantly lower than PRE levels ($p = 0.001$). Also, a significant reduction in self-ratings of emotional arousal from DURING (immediately after the end of the activity performance) to POST (after a 2-min rest period) was observed in both uPMR ($p = 0.05$) and CA ($p < 0.001$). Regarding emotional valence, there was no significant main effect of condition, $F(1, 32) = 0.136, p = 0.715$. There was a significant interaction effect for condition × time, $F(1.80, 57.84) = 4.28, p = 0.022$. Pairwise comparisons revealed that PRE valence levels differed between conditions ($p = 0.049$). No significant difference in valence levels was found in DURING ($p = 0.823$) and POST ($p = 0.085$) levels. Given the lower valence baseline (PRE) in uPMR, results suggest that uPMR had a more positive effect on valence. In other words, uPMR may have brought participants to a higher valence state from PRE to DURING ($p = 0.568$) while CA worked in the opposite direction with negative effects on valence from PRE to DURING ($p = 0.630$) and DURING to POST ($p = 0.034$). The CA was a non-stimulating activity that consisted of watching a video and remaining passive. This could be the reason participant levels of arousal decreased and they switched to a more negative valence state (i.e., increased boredom).
While changes from baseline (PRE) to immediately after activity performance (DURING) did suggest higher arousal and more positive affect in the uPMR condition, these changes were not significant. Thus, our findings differ from those of Schienle and Unger (2021) in which university students reported significantly lower arousal and more positive affect after daily 10-
min PMR exercises. A possible explanation is that participant familiarity in performing PMR and activity duration are important factors related to generating changes in affective states.

**State anxiety (InSTAI-S):** There was no significant main effect of condition, $F(1, 32) = 0.109$, $p = 0.743$; no interaction effect for condition $\times$ time $F(1.83, 58.55) = 0.832$, $p = 0.431$. Pairwise comparisons revealed that only uPMR showed a significant reduction in self-ratings of state anxiety from PRE to POST ($p = 0.036$). Significant reductions in state anxiety from DURING to POST were found in both uPMR ($p = 0.05$) and CA ($p = 0.01$). Thus, levels of anxiety did not differ from baseline to the measure immediately after uPMR performance (DURING) but only from baseline to the measure taken after a 2-min rest period (POST). These findings differ partially from those of Pawlow & Jones (2005) in which they compared a single, 20-min PMR practice to a control and found significantly reduced post-activity self-ratings of state anxiety. Thus, it seems the duration of practice is an important factor contributing to the psychological effects of PMR.

In sum, H2 was only partially supported, as there was a significant decrease in levels of anxiety compared to baseline 2 minutes after uPMR training. However, uPMR did not modify valence and it reduced levels of arousal similarly to CA as measured by AS. Also, it should be noted that the level of anxiety measured 2 minutes after and not immediately after uPMR training was lower than baseline while activation of PNS (as indicated by changes in InRMSSD) was observed during the training and not after. An explanation for this physiological and psychological desynchronization could be that the emotional experience of anxiety seems to be dependent on both autonomic nervous system feedback and the cognitive interpretation of the stimulus that induced these bodily signals (Nummenmaa et al., 2014). In other words, emotional states arise from physiological changes and interoceptive signaling occurs before there is emotional awareness (Critchley & Garfinkel, 2017).

<table>
<thead>
<tr>
<th></th>
<th>Arousal</th>
<th>Valence</th>
<th>InSTAI-S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>uPMR</td>
<td>CA</td>
<td>uPMR</td>
</tr>
<tr>
<td>Pre</td>
<td>2.74 ± 0.82</td>
<td>2.81 ± 0.75</td>
<td>3.44 ± 0.57</td>
</tr>
<tr>
<td>During</td>
<td>2.71 ± 0.76</td>
<td>2.95 ± 0.76</td>
<td>3.53 ± 0.53</td>
</tr>
<tr>
<td>Post</td>
<td>2.48 ± 0.76</td>
<td>2.47 ± 0.88</td>
<td>3.46 ± 0.49</td>
</tr>
</tbody>
</table>

### 3.3 H3: Student Opinions Collected from the Post-Trial Survey

Descriptive statistics can be found in Table 2. Across items related to the acceptability of the activity for use in formal educational settings (items 5, 6, 9-11), ratings for uPMR were all above 4 (out of 7) and significantly higher than CA suggesting that participants would accept uPMR for use in formal educational settings. Furthermore, participants found uPMR easy to perform with 2 minutes being rated highly as an appropriate duration. Ratings for items 2-4 (effects on arousal and mood) are in line with the psychophysiological results previously reported. Thus, H3 is confirmed as participants rated uPMR as more favorable and acceptable for use in formal educational settings than CA. A previous study by O’Donnell & Dunlap (2019) also found high acceptance rates of digital audio guided PMR interventions among schoolteachers. Thus, uPMR has the potential to be an accepted activity by both students and teachers. Although, a follow up study with university educators would help confirm this.

In response to the open question, 11 (27.5%) of participants commented that they had difficulty relaxing during the activity. Thematic coding (Gibbs, 2007) of the reasons given generated 4 codes. Responses referred to a lack of comfort with the activity (2), lack of comfort with the environment (2), problems with the nature of the activity such as it being too long, too isolating, and not relaxing to contract muscles (3), and ruminating while or having general discomfort with relaxation activities which prevents relaxation (4). As an example of the latter,
one participant wrote, “It has always been difficult for me this (sic) relaxing activities because I don’t feel comfortable.” The reason for the latter responses could be that initial performances of basic types of relaxation activities such as PMR or mindful breathing require effort and attention. This can make the activities less relaxing and/or even increase stress levels. With practice less effort and attention are required, and greater relaxation benefits are realized. This is known as the paradox in relaxation training (Heide & Borkovec, 1983) and is more prevalent in people with higher basal levels of anxiety or anxiety disorders (Kim & Newman, 2019). This paradox may also be a factor that influenced the self-reports for affect and anxiety.

<table>
<thead>
<tr>
<th></th>
<th>uPMR</th>
<th>CA</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The activity is easy for me to perform.</td>
<td>6.25 (1.30)</td>
<td>5.58 (1.57)</td>
<td>0.043</td>
</tr>
<tr>
<td>2. The activity calms me down.</td>
<td>5.73 (1.55)</td>
<td>3.68 (1.90)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>3. The activity energizes me.</td>
<td>1.93 (1.29)</td>
<td>2.38 (1.35)</td>
<td>0.098</td>
</tr>
<tr>
<td>4. The activity improves my mood.</td>
<td>4.03 (1.39)</td>
<td>3.43 (1.63)</td>
<td>0.041</td>
</tr>
<tr>
<td>5. The duration of the activity is appropriate.</td>
<td>5.80 (1.22)</td>
<td>4.98 (1.58)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>6. I would like to do this activity as part of a class.</td>
<td>4.53 (1.83)</td>
<td>3.25 (2.06)</td>
<td>0.003</td>
</tr>
<tr>
<td>7. I would feel embarrassed performing this activity in a class.</td>
<td>2.10 (1.75)</td>
<td>1.68 (1.39)</td>
<td>0.184</td>
</tr>
<tr>
<td>8. This activity would help me be more open with others.</td>
<td>2.70 (1.22)</td>
<td>1.95 (1.20)</td>
<td>0.001</td>
</tr>
<tr>
<td>9. This is an appropriate break activity to use in an online class.</td>
<td>4.20 (2.10)</td>
<td>2.78 (1.76)</td>
<td>0.002</td>
</tr>
<tr>
<td>10. ... in an in-person class.</td>
<td>4.58 (1.88)</td>
<td>3.28 (1.77)</td>
<td>0.004</td>
</tr>
<tr>
<td>11. ... when I study on my own.</td>
<td>4.98 (1.94)</td>
<td>3.20 (1.73)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

4. Limitations and Future Directions

There are several limitations to this study that can be used as suggestions for future work. (1) The study was conducted in a controlled, experimental setting with individual participants rather than in everyday settings of participants which limits the generalizability of results to real-life settings. (2) Despite the sample size being similar to other physiological studies of self-regulation activities (You et al., 2022; Nickel et al., 2005; Pawlow & Jones, 2005), a larger sample should be considered to account for potential technical issues causing the loss of data. (3) While CA was intended to be a neutral activity and was used as such by You et al. (2022), this study found that it produced significant effects on affect and PNS activity. Different control activities that are more active-neutral (Smith & Norman 2017) or based on leaving free time to the participant (Crocker & Grozelle, 1991) should be considered.

Beyond addressing the limitations, further investigations should explore whether the effects of uPMR differ by gender, ethnicity, and among disadvantaged students. Additionally, to make uPMR more inclusive, future work is needed to identify how to support those not able to relax when performing the activity such as by altering guidance offered or providing additional scaffolding. Finally, microlearning which focuses on delivering small chunks of skills-based, just-in-time learning (Zhang & West, 2020) should be explored as a way of developing student self-regulation competences. For example, ultra-brief self-regulation activities can be taught and practiced in stressful academic situations such as prior to exams and presentations but form part of a longer, cumulative training course on self-regulation skills.

5. Conclusion

Progressive muscular relaxation is an evidence-based intervention for reducing physical symptoms of anxiety and improving well-being. The current study findings suggest that an ultra-brief, video guided PMR activity possesses promising characteristics for use in situated teaching and practice of self-regulatory techniques aimed at reducing physiological arousal (a symptom of anxiety) and stress on an acute basis in university students. Overall, this work aims to support students in developing critical life skills (resilience, self-management, meta-cognition) by identifying evidence-based self-regulation activities appropriate for formal
educational settings and supporting educator use of these activities. The latter can be achieved through technologies such as ClassMood App that help teachers create classroom opportunities for guiding students through evidence-based self-regulation practices.

Acknowledgements

This work has been partially funded by Pompeu Fabra University’s Planetary Well-being Initiative (PLAWB00321, BEAM), Spanish Ministry of Science and Innovation (PID2020-112584RB-C33 granted by MICINN/AEI/10.13039/501100011033), Department of Research and Universities of the Government of Catalonia (SGR 00930), and European Union-Next Generation EU.

References


The Development and Evaluation of the Platform for Online High-Level Cooperative Games

Feng-Lung LIUa, Geng-De HONGb*, Ju-Ling SHIHc & George Ghinead

abcGraduate Institute of Network Learning Technology, National Central University, Taiwan
dDepartment of Computer Science, Brunel University London, UK
*hgengde@gmail.com

Abstract: The design of educational games plays an increasingly important role in modern education. By combining learning with games, it can attract students' attention and making learning more interesting. This study describes the development of a platform designed specifically for the online high-level cooperative game in which students are encouraged to achieve their own goals as well as common goals to reach win-win situation. Technology Acceptance Model (TAM) was used to analyze students' overall perceptions of the platform. The results show that the design of the game platform had high perceived usefulness and perceived ease of use, and players also showed positive attitudes and usage intentions. Therefore, a good game platform should have interactivity, educational value, problem-solving settings, and usability to provide the best gaming experience and have a positive impact on learning.

Keywords: Game-Based Learning, Gaming Platform, System Development, High-Level Cooperative Games, Technology Acceptance Model

1. Introduction

1.1 Research goal

In traditional learning methods, students usually need to spend a lot of time and effort learning knowledge and skills, which can lead to learning fatigue and loss of interest. Today, some games are gradually integrating educational resources and elements to enhance students' learning effectiveness and interest. According to Petri and von Wangenheim (2017), incorporating interactive, exploratory, and problem-solving elements into games can enhance learning outcomes by boosting students' engagement, and motivation; resulting in better learning outcomes.

The design of educational game platforms can contribute to students' experience of the pleasure of success and sense of achievement in games. This, in turn, can enhance their learning motivation and interest. Therefore, this study involved the development of an online high-level cooperative game platform. The platform provides various game modules for resource sharing, trading, bargaining, and issue discussions among players. After the game, reflections and feedbacks were received using the TAM questionnaire to investigate students' learning attitudes and technology acceptance. Evaluation results help system designers to develop more effective online game systems and enable continuous improvements of the game system.

1.2 Research questions

This study aims to understand students' experiences, acceptance, and willingness to use the online high-level cooperative game system using the Technology Acceptance Model (TAM). The research question of this study is to investigate students' perceived usefulness, perceived ease of use, attitudes toward use, willingness to use, and overall feedback experience of using the game system.
2. Related work:

2.1 Game-Based Learning

Digital games have emerged as effective learning tools that can assist students in acquiring knowledge while playing. Gee (2003) argues that digital games can enhance students’ problem-solving, analytical, and creative abilities, and even help them develop skills beyond the classroom. Originally conceived as entertainment, digital games have become widely used as a learning tool, thanks to the prevalence of 3C technology products. Digital games provide an entertaining environment that increases student engagement and motivation (Annetta et al., 2009). Laine and Lindberg (2020) have demonstrated that digital games possess powerful motivational factors that can spark learners' initiative. By providing an engaging and interactive setting that challenges and entertains, digital games facilitate learning and cultivate students' problem-solving, analytical, and creative skills. Given their significant motivational benefits, digital games have become a popular tool for learning and are anticipated to remain a crucial component of education.

2.2 Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM) was initially proposed by Davis, Bagozzi, and Warshaw (1989) to investigate the factors influencing individuals' willingness to use new technology. TAM comprises of two fundamental elements that substantially influence a user's inclination towards adopting new technology and shaping their behavioral attitudes. Perceived usefulness pertains to the extent to which users deem that utilizing a system can boost work efficiency. Perceived ease of use denotes the user's subjective perception of the system's complexity or simplicity, which is closely associated with their level of information technology proficiency. In order to further explore the acceptance of other technologies, it is necessary to consider the attitude toward using, which includes the user's perception of factors such as interface environment and system quality, and how they influence the user's behavior. Behavioral intention is influenced by multiple variables and ultimately reflects the user's actual operational behavior. The TAM has been widely used in research on user acceptance of technology and is considered an established model for understanding technology acceptance (Gefen, Karahanna, & Straub, 2003). Examples of its application include e-learning systems (Nga et al., 2007), online courses (Landry et al., 2006), websites (Koufaris, 2002), and others. In addition, TAM can be further expanded by incorporating external variables and combining with other theories. Different external variables may be related to system, user, and task characteristics, which can be investigated in research.

3. Online High-Level Cooperative Games Platform Design

This study combines game-based learning and issue-based learning to develop a high-level collaborative game-based learning system. The system incorporates historical crisis events, and through game mechanics, players can solve game problems through trading and negotiation. The system design is based on the Great Voyage game developed by Hong, Shih, and Lu (2022), allowing players to negotiate and divide tasks among groups to resolve historical crises. Students are divided into five groups to role-play European countries: England, France, Portugal, Spain, and the Netherlands, each with their own tasks. In the game, players go through four stages of gaming process in every round. The first stage is internal affairs, in which countries generate strategic plans for crisis resolution. The second stage is diplomacy, every country negotiates to resolve conflicts. The third stage is declaration, in which country representatives state their stands and propose solutions to the crisis event. The final stage is settlement, in which the Pope assesses the situation of each country and makes judgments. The game ends when all events are effectively resolved.

In respond to this game process, this study designed four functional modules in the system. Provide players with a good gaming experience, this system with simple user
interface, clear game information, and obvious text and icons to guide students to complete game tasks smoothly and foster their learning motivation. The design of each module is introduced and explained below:

(1) Information Model (Figure 1):

The Information Model interface not only displays relevant information about the player’s own country in the game, but also allows for comparison through the intelligence information screen after each crisis is resolved. By comparing the total assets and trends shown in the initial information and current status, players can obtain valuable information about their own country and make wiser decisions to achieve victory. This transparency of information can help players better understand the situation of their own country, optimize their resource allocation, and effectively respond to crisis challenges. This feature enhances the fun and challenge of the game, while also giving players a sense of self-improvement.

(2) Negotiation Model (Figure 2):

In the Negotiation Model interface, players can freely choose different strategy behaviors, moods, message recipients, and discussion topics to better meet their needs and negotiation goals. This allows players to participate more comprehensively in negotiation topics and dialogue processes, fosters cooperation and communication, and increases player involvement in resolving crises. During the negotiation process, text messages can be displayed in real-time, allowing players to conveniently understand the negotiation process and results; the different colors of the dialogue box can distinguish messages from different countries, further enhancing the visual effects and user experience. This interface design enhances the freedom and fun of player participation in negotiation and dialogue, while also strengthening the realism and interactivity of the negotiation process, enabling players to better experience and learn negotiation skills and strategies.

(3) Trading Model (Figure 3):

The Trading Model provides multiple options and trading methods, allowing players to easily select different trading modes, resource items, commodity categories, trading partners, trading items, and trading commodities, thereby increasing the flexibility and options of resource trading for players. This interface design not only enhances the freedom and fun of players’ participation in resource trading but also strengthens their decision-making ability and strategic thinking in resource trading. Players can flexibly choose different trading methods and commodity categories based on their own needs and goals, in order to achieve maximum economic benefits and game objectives, helping players to understand the essence and rules of resource trading.

(4) Auction Mode (Figure 4):

The Auction Model interface provides players to participate freely in product auctions and transactions, and to list their own products, conduct auctions, and choose to buy spices and other goods, as well as query transaction records after completion. Through these operations, players can effectively understand the general process of the transaction and thus enhance their understanding and comprehension of the transaction. At the same time, it can also help players improve their trading skills and strategic abilities, thereby gaining more economic benefits in the game. More importantly, this platform allows players to gain more freedom and fun in participating in auctions and transactions, while also enhancing their understanding and mastery of trading skills and strategies in the game.
4. Research design

4.1 Research Framework

This study recruited fifteen Taiwanese graduate players to participate, who were randomly assigned into five groups. The research procedure, commenced with an explanation of the game rules and the system's operation before the game's commencement. The gaming activities were then documented in the game system. During the game (Figure 5), all players engage in online domestic discussions and diplomatic negotiations to solve problems and achieve the assigned tasks. The system supports students to tackle multiple crisis events which are supported by the system with various functional modules. Subsequently, upon completion of the game, the students were guided to complete a feedback questionnaire concerning the game system.

4.2 Research Tools

The questionnaire design of this study is based on the TAM proposed by Davis et al., In the TAM, the external variables of perceived usefulness and perceived ease of use are explored. This study focuses on "computer self-efficacy" as the main factor to examine whether players have the ability to use information systems to complete tasks. Therefore, the game content was taken into consideration; appropriate design and adjustments were made with the results of a questionnaire. The questionnaire was divided into five dimensions, including computer self-efficacy, perceived usefulness, perceived ease of use, attitude toward using, and behavioral intention to use. Each dimension includes five questions to understand players' acceptance and satisfaction with the system platform. Further evaluation of the system performance was done in terms of technology acceptance.

5. Results

Players' perceptions using online high-level cooperative games platform was analyzed with TAM (Table 1). According to the data results, it can be observed that the average value of each dimension is higher than the medium score of 3.00 of the five-point likert scale. These high ratings reflect the students' positive attitude towards the game, and their willingness to using it in the future. They believe that the game can make it easier for them to learn.
Table 1. TAM evaluation to the game platform

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Efficacy for Computer</td>
<td>3.73</td>
<td>1.02</td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>3.76</td>
<td>0.73</td>
</tr>
<tr>
<td>Perceived Ease of Use</td>
<td>3.32</td>
<td>0.84</td>
</tr>
<tr>
<td>Attitude towards Use</td>
<td>3.92</td>
<td>0.71</td>
</tr>
<tr>
<td>Behavioral intention</td>
<td>3.81</td>
<td>0.67</td>
</tr>
</tbody>
</table>

In the players' self-efficacy for computer dimension, the overall mean score is 3.73 (SD=1.02). The various sub-dimensions have average scores ranging from 4.33 to 3.13. Among them, the highest score is for Q1 "I often use computer devices to aid in my learning" (m=4.33), indicating that players who frequently use computer devices to assist learning. However, the lower scores are seen in Q2 "I often use computer devices to play online games" (m=3.73), suggesting that players use computer devices less often to play online games. This shows that players have a higher acceptance of using computers to play online games, indicating that they are better able to cope with the challenges and problems in various computer games, and have confidence and ability in using the platform.

In the players' perceived usefulness for computer dimension, the overall mean score is 3.6 (SD=0.73), with the mean values of sub-dimensions ranging from 3.93 to 3.53. The values indicate that most players have a high level of agreement with the "high-level collaborative online game platform". The highest score is seen in Q10 "Overall, using a 'high-level collaborative online game platform' is beneficial for me" (m=3.93), suggesting that players find the platform highly beneficial. However, in Q8 "The 'high-level collaborative online game platform' saves me a lot of time and effort when I am learning" (m=3.53), the lower score indicates that players perceive the platform to be limited in terms of time-saving and effort-saving during learning. The overall result shows that players still perceive the game system as assisting them in achieving their gaming or learning goals.

In the players' perceived ease of use dimension, the overall mean score is 3.32 (SD=0.84), and the mean values of each sub-dimension range from 3.47 to 3.20. The highest score is seen in Q15 "Overall, I find the 'high-level collaborative online game platform' to be easy to use" (m=3.47), indicating that players find the platform easy to operate, with clear and concise functions that allow for easy learning and quicker mastery of the system's features. However, the lower score is observed in Q13 "I am proficient in using the 'high-level collaborative online game platform'" (m=3.20), suggesting that players may find it difficult to become familiar with the platform's use due to the complexity of the game's mechanisms. Overall, players perceive the game platform as easy to use, with clear and concise functions that enable them to learn and master the system quickly.

In the players' attitude towards use dimension, the overall mean score is 3.92 (SD=0.71), and the average scores of each sub-dimension range from 4.07 to 3.67, indicating that players are highly satisfied with the overall experience of the system. The highest scores are observed in Q18 "I think that the 'high-level collaborative online game platform' is an attractive learning tool" and Q20 "Overall, I think that using the 'high-level collaborative online game platform' is a good idea" (m=4.07), suggesting that players believe the game promotes learning and cooperation. Players find the platform easy to use and appreciate its practicality, which helps them achieve game objectives and effective learning.

In the players' behavioral intention to use dimension, the overall average score is 3.81 (SD=0.67), with average scores ranging from 3.87 to 3.80 in each sub-dimension, indicating a high willingness to use the system among the participants. Among them, in Q24 "I expect to continue using the 'high-level collaborative online game platform' in the future" (m=3.87) is the highest score for the expectation that they will continue to use. This indicates that learners have a high willingness to use the system, are willing to spend time and effort using the game for learning, and are satisfied with its good experience and functionality, providing an excellent learning environment.
6. Conclusion

This study developed an online gaming platform with high-level cooperation, providing various system functional modules to assist players in effective communication, collaboration, conflict resolution, issue discussion, and reflection during the game. After the game, the TAM model was used to evaluate the game, and the questionnaire design was appropriately adjusted based on the game content to understand players' technology acceptance and satisfaction with the system platform. The results showed that players had positive evaluations of the various system functional modules, which helped them achieve their learning goals and provided a good gaming experience. It also stimulated their learning interest and motivation, met their gaming needs and expectations.

According to research findings, a good gaming platform should have interactivity, educational value, problem setting, and ease of use to provide players with the best gaming experience and play a positive role in learning. Interactivity promotes player interaction and communication with other players and enhances team collaboration awareness. Setting challenges and problems in the game allows students to learn through problem-solving and further stimulates their learning motivation and interest. Ease of use is also essential; simple operation and a good user experience allow players to easily immerse themselves in the game. By designing games in this way, it can not only improve students' problem-solving skills and creativity but also promote communication and interaction among students, enhancing their social skills and teamwork spirit.

The ongoing system development seeks to tailor the platform for global educational endeavors, with the goal of investigating how participants from diverse cultural backgrounds interpret and engage with the various in-game roles. Anticipating cross-cultural comparisons is part of this endeavor.

Acknowledgements

This study is supported in part by the National Science and Technology Council (previously known as Ministry of Science and Technology) of Taiwan, under MOST 108-2511-H-008-016-MY4 and 112-2918-I-008-003

References


A Thematic Analysis Exploring Flexibility in Programming-based Mathematical Problem Solving

Huiyan YEa, Oi-Lam NGa* & Zhihao CUIa

aFaculty of Education, The Chinese University of Hong Kong, China
* oilamn@cuhk.edu.hk

Abstract: Research on computational thinking (CT) have received significant attention in recent years, but how CT and mathematical thinking (MT) work together to influence problem solver’s computational strategies and competency has not been illuminated. Taking the method of thematic analysis combined with a conceptual framework of flexibility in programming-based mathematical problem solving, this study aims to shed light on the interplay between CT and MT. We identify two kinds of flexibility when a problem solver engages in programming-based mathematical problem-solving: translation flexibility and integrated flexibility. We hope that this research will contribute to a deeper understanding of the interplay of CT and MT as a basis for conceptualizing an interdisciplinary form of STEM education and thinking on task design, pedagogy, and assessment approaches of CT-based mathematics instruction.

Keywords: flexibility, computational thinking, mathematical thinking, interplay

1. Introduction

Computational thinking (CT) is regarded as a fundamental skill required in daily life, and a kind of analytical thinking that shares common characteristics with mathematical thinking (MT) in many ways (Wing, 2011). Thus, researchers have investigated ways to integrate CT into mathematics teaching and learning (e.g., Ng & Cui, 2021). However, one of the key challenges is to understand how CT and MT work together to influence students’ strategies when engaging in mathematical problem-solving in computational contexts. Previous studies have identified discrepancies between CT and MT, which make mathematical problem-solving in computational contexts more challenging for students (Cui & Ng, 2021). Therefore, understanding the complex relationship of CT and MT is vitally important if CT is to be meaningfully integrated into mathematics education.

While the breadth of research has advanced the empirical basis of understanding what CT-based mathematics instruction entails, very little is currently known about the processes of how CT and MT specifically work together in this context. Meanwhile, noting the diversity of solution strategies in the CT-based mathematical problem-solving, research on flexibility in mathematical problem-solving may provide possible directions for our attempts to explore the interplay between CT and MT. Flexibility refers to “the ability to change according to particular circumstances” (Star & Seifert, 2006, p. 281), which relies on the problem solver’s understanding of the knowledge relevant to the problem situation. If the problem context is extended to a CT-based mathematical problem, not only would multiple programming solution strategies be possible, but the problem solvers would also require knowledge of programming and mathematics, as well as to adapt this knowledge according to the programming environment used and mathematical ideas involved respectively within the problem context. In this study, we are interested in delving into this very interplay between CT and MT from perspective of flexibility. Therefore, the study is guided by the research question: What types of flexibility are exhibited during CT-based mathematical problem-solving?
2. Theoretical Background

2.1 Computational Thinking and Mathematical Thinking

In this study, we define CT as “the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent” (Wing, 2011, p. 20), and MT not as “thinking about the subject matter of mathematics but a style of thinking that is a function of particular operations, processes, and dynamics recognizable mathematical” (Burton, 1984, p. 35). In other words, both CT and MT refer to the cognitive structures that govern an individual’s programming and mathematical activities respectively, as well as the anticipation of the results of those activities. CT is thought to share a similar approach to MT in problem solving (Wing, 2011). Researchers also attempted to find out some of the empirical connections between CT and MT, with some reckoning that CT and MT are mutually supportive or have overlapping features. For example, Ng and Cui (2021) argued that mathematical exploration can serve as a site that allows students to engage in CT in meaningful ways. On the other hand, the presence of CT and MT may occur in mismatching ways. As Rich and colleagues (2018) point out, mathematics focuses on students’ flexible arithmetic operations and arithmetic strategies, while CT focuses on sequence so that the computer can recognize and execute. While these studies point to an inter-relationship between CT and MT (e.g., CT support MT, or vice versa, or the two are constraining), the conclusions are relatively broad and lack a detailed perspective on how such facilitative or hindering relationships of CT and MT arise and develop. That is, there requires more evidence to capture the presence or expression of CT and MT in a clear way and how they affect students’ thinking processes when they engage in CT-based mathematical problem solving.

Flexibility is considered to be a key competency in problem solving. Even if research focusing on flexibility has been studied extensively in different domains of mathematics, there has been little research on interdisciplinary flexibility, that is, how flexibility manifest across different subject domains. In particular, the flexibility of working across CT and MT in the context of CT-based mathematical problem-solving not yet been empirically studied and analysed. In this study, we adopt flexibility as “knowledge of multiple strategies and the ability to select the most appropriate strategy for a given problem and problem-solving circumstances” (Star et al., 2022, p. 2) so that we can observe how flexibility are demonstrated through various tasks in our designed CT-based mathematical problems. This investigation can potentially refine our understanding of how problem solvers make connections between computational and mathematical concepts. In particular, we pay attention to the tasks in which problem solvers are situated to translate between mathematical and programming language or integrate mathematics concepts and computational concepts while making meaningful connections of CT and MT. When analysing these tasks, we recognise that programming itself is a tool-mediated activity; therefore, the instrumentation of different programming tools would necessitate different kinds of CT-MT translations and connections to occur. For this reason, we choose to address one specific programming language, Scratch.

3. Methodology

3.1 Context and Participant

This study is part of a larger design-based research addressing the research objectives of designing a series of tasks for supporting computationally enhanced mathematical problem-solving in school mathematics classrooms, and of studying student learning
trajectories in terms of developing their CT, MT, and problem-solving competence in the designed tasks. The participants in this study were from two programming-based mathematical problem-solving programs. One was a student program, which was design-based research (DBR) undertaken with three iterative cycles of implementations with seventh to ninth grade (age 12 to 15) students (n = 95) from three secondary schools in Hong Kong. Another one was a teacher program, which adopted task-based semi-structured interview to collect data. Eight in-service mathematics teachers from different schools in Mainland China, participated in a five-week online professional development workshop in Scratch programming. The two programs were conducted independently upon obtaining participants consent, with the student program being undertaken in 2021 and the teacher program being conducted in 2022 via Zoom.

In this study, we adopt the method of thematic analysis to further interpret the translation/connection involved between CT and MT in the programming solutions to the chosen tasks through the lens of flexibility. As the purpose of this study was to explore the flexibility of learners’ strategies in solving programming-based mathematics problems, although participants included two different groups (i.e., students and in-service mathematics teachers), we believe our data still appropriate to answer the RQs, since the programming solutions, as opposed to the participants (or problem solvers) were the focus of this study.

3.2 Data Collection and Analysis

Each program contained an introduction game and several tasks related to four domains: number, algebra, geometry, and data handling. A total of 18 tasks were studied in response to the research question posed, in which students program contained 14 tasks (1 introduction game, 5 in number and algebra, 4 in geometry, 4 in data handling) and teachers program contained 4 tasks with one task in each topic. All tasks are programming-based mathematical problems and are all open-ended problems with no fixed coding strategy. For example, “Please write a code project that can be used to verify whether an entered number is prime or not.” Coding-related solutions are all thought out and collaborated by participants (students or teachers). The researchers will observe participants’ problem-solving process and guide the participants to reflect through task-based interviews at the appropriate time, which aims to understand what they think and to help them solve the difficulties they encounter during the programming process. Therefore, our data include (a) video and voice recording of interactions and interview between researchers and participants; (b) collection of field notes, screen capturing, developed artefacts and documents in both programs, which help us to gain a deep understanding of the research questions.

As a qualitative analysis method to identify, analyze and report patterns of data (Braun & Clarke, 2006), thematic analysis is regarded as an effective approach for identifying similarities and differences, yielding unforeseen insights and summarizing key features of a large data set (Nowell et al., 2017). Therefore, we adopted thematic analysis with the aim to identify how participants exhibited flexibility in their programming solutions for mathematical problems. Taking our designed tasks and participants programming solutions as the unit of analysis, we conducted latent thematic analysis, which developing the themes necessitates interpretive effort and the resulting analysis is not just descriptive, but is already theorized (Braun & Clarke, 2006). Our analysis was implemented following the six steps of doing thematic analysis outlined by Braun and Clarke (2006).

4. Results

By analyzing and generalizing participants’ problem-solving strategies in solving CT-based mathematical problems, the results of the thematic analysis yielded two main kinds of flexibility (i.e., translation flexibility and integrated flexibility) and six sub-categories. Translational flexibility refers to a cross-disciplinary, cross-problem context capability that involves translation between mathematical and programming language, with the two processes being independent. In contrast, integrated flexibility is an overall capability similar
to flexibility in a mathematical problem context (e.g., to solve an algebra equation; Star & Seifert, 2006). It is the flexibility to solve problems in a programming environment that involves the use of knowledge from both disciplines, with the MT and CT realizations being inseparable as a whole. Due to the length of the article, we will only cover integrated flexibility in detail with a geometry example.

One of the most evident examples of integrated flexibility is the setup of subroutines, often known as creating *My Blocks* in Scratch. As an important CT concept, subroutines can be divided into two types. One is the parameterless subroutine, which calls an entire piece of code directly and repeatedly as a whole, without altering it before or after the subroutine function is invoked, but rather given it a separate name for wrapping (Figure 1a-b). This type of subroutine does not require programmer to make any adjustments to the codes. Instead of executing the original code directly, another kind of subroutine with parameters requires programmers to comprehend the section of codes in its entirety, identify the key attribute and numerical relationships between them, and then modify the codes by abstracting parameters to be a subroutine (Figure 1c). In this section, we will focus on subroutines with parameters and the related integrated flexibility will be examined through a geometry drawing task.

As we know, to draw a regular polygon in paper-and-pencil context, we first need to know the polygon’s side length, the degree of interior angle, then to draw it with a ruler and a protractor. While to achieve this mathematical process in Scratch, the Pen function is used to record the trajectory of a Sprite. For example, to draw a regular pentagon with side length of 50, the pen needs to [move 50 steps, turn clockwise 72 degrees] then repeat this process five times (Figure 1a). The code at this point does not yet make use of subroutine, but if programmers try to use *My Block* to simplify the programming process or design a program that can change the size and shape of the polygon at any time, mathematical consideration is required for setting parameters in the subroutine (Figure 1c-d).

It is worth noting that the process of changing the original codes (Figure 1a) to a piece of code with subroutine (Figure 1c-d; Figure 2a-b) requires the involvement of both MT and CT. The interplay between CT and MT is evident in this integration, particularly in the different strategies from Figure 1c-d to Figure 2a-b. For example, programmers can abstract a parameter “side length” in the *My Block* and replace all the numbers or variables in the source code that represent the length of the sides to create a subroutine served to control the size of a pentagon (Figure 1c). Here, we can see that such a change is mediated by both CT and MT because one may have to figure out which variable represent the side length that can be a parameter in the subroutine. CT support programmers to create the subroutine,
while MT helps determine the first parameter in the subroutine, i.e., the size of the polygon is determined by the length of the sides. However, when the programmer went on to optimize the subroutine so that it can also change the shape of the polygon, it does not mean that mathematical property “different polygons have different interior angles” makes it possible to add a second parameter “exterior angle” directly to the My Block in Figure 1c. Otherwise, even if the second parameter “exterior angle” is added to the My Block (Figure 1d), it will not be possible to draw any other regular polygon than a regular pentagon, because only with the parameter “polygon 50 72” will the matching angle and number of sides can form an enclosed polygon. This is because the number of sides and the degree of interior angles determine the shape of the polygon, while the subroutine in Figure 1c only contains codes of [move 50 steps, turn clockwise 72 degrees] five times which does not have the function to change number of sides that follow the change of degree of exterior angle. The key to solve such problems is the effective coordination of CT and MT.

Furthermore, although the subroutine in Figure 2a with three parameters (i.e., “side_length”, “exterior_angle”, “number_of_sides”) can change the size and shape of the regular polygon, there actually exists a specific quantitative relationship between the degree of exterior angle and the number of sides, i.e., degree of exterior angle = 360/number of sides, so the My Block in Figure 2a can be optimised to a subroutine with only two parameters affecting the shape (number_of_sides) and size (side_length) of the polygon in Figure 2b. This reveals the synergy between CT and MT in the process of setting up subroutines with parameters. Thus, we can observe that during the process of setting subroutines, the key concepts of CT (subroutines, loops) and MT (mathematical properties) constrain and facilitate each other: When problem solvers want to set up flexible subroutines that can change the shape and size of regular polygons, CT concepts such as subroutines and loops play a key role in supporting the adaptation of the program, while quantitative relationships in mathematics involved in MT responsible for decision making on parameters setting in the subroutine, the choice of using which CT concepts and how they are used. That is, the number of sides determine the number of loops; the parameters of the subroutines are the attributes that control size and shape—the length and the number of sides, respectively. It is in the combined effect of CT and MT that provide opportunity for programmers to abstract the key parameter that determines the shape of a regular polygon. We have named such an integrative programming process as “integrated flexibility” in which CT and MT co-exist and support each other.

5. Discussion

Overall, translation flexibility is a cross-disciplinary, cross-problem context competence, while integrated flexibility is a holistic capability similar to flexibility in mathematical problem
contexts (Star et al., 2022; Star & Seifert, 2006), that is, the general flexibility to solve problems in a programming environment that includes knowledge from both mathematics and programming disciplines. It involves a complex interaction between MT and CT which differ from the one-way conversion of translation flexibility. In general, integrated flexibility is more often found in the process of selecting a CT technique when the problem solver has a clear mathematical solution. It is vital to take into account which mathematical approach and which CT technique are compatible. If the two do not match, adjustments must be made to the mutual circumstance in order to develop an effective programming solution.

We believe such a conceptual framework has potential for understanding the mutual cognitive development in CT and MT, including how mathematical concepts co-emerge with programming processes in a problem-solving context, and the role of CT in such processes. We understand that because programming is a tool-mediated activity, different programming tools will be instrumented differently, necessitating various CT-MT translations and interactions. It is hoped that this research will contribute to a deeper understanding of the interplay of CT and MT as a basis for conceptualizing an interdisciplinary form of STEM education and thinking on task design, pedagogy, and assessment approaches of CT-based mathematics instruction. Moreover, the findings will offer theoretical contributions that inform researchers and mathematics educators in terms of student learning characteristics during CT-based mathematics instruction. Finally, we have noticed that our data from two different participant groups exhibit some comparable features of flexibility, which may point to a certain universality of problem solvers’ flexibility in programming-based mathematical problems. As a concluding remark of the paper, we pose several questions that arise from our exploration—Given the existence of multiple solutions, are the prevailing programming solutions simpler or more complex to solve the problem? Why are these approaches more common? What factors are relevant to them? How are the problem solvers’ flexibility to use different types of translations affected by their CT and MT? We call for more research to consider the different kinds of flexibility demonstrated in CT-based mathematical problem solving to sketch the big picture of how CT and MT emerge in complementary or mismatching ways.

References


Epistemic Network Analysis to assess collaborative engagement in Knowledge Building discourse

Aloysius Ong*, Chew Lee Teoa, Alwyn Vwen Yen Leea & Guangji Yuan*

aNational Institute of Education, Nanyang Technological University, Singapore
*aloysius.ong@nie.edu.sg

Abstract: Knowledge Building (KB) is an established learning sciences theory that seeks to promote innovative ideas and idea improvement among students via collaborative engagement in productive discourse. KB discourse supports students to make constructive discourse moves such as questioning, explaining with evidence, adding new information and so on, to advance the collective inquiry. However, current understanding on KB discourse remains limited to students’ online participation. Although small group discussion is a common practice, there is little understanding on the role of verbal discussions to support KB discourse. This paper attempts to address this line of inquiry by assessing student engagement in KB discourse supported by both online and verbal discussions. Data is retrieved from a group of six students in a Grade 6 Social Studies class. The group participated in a 2.5hr lesson designed with opportunities for discussions on the Knowledge Forum (online) and in small groups (verbal). Group talk was transcribed, and Knowledge Forum notes were coded for its semantic level of contribution, with the codes being analysed for weighted connections using Epistemic Network Analysis (ENA). The ENA analysis revealed clear differences in both group and individual engagement between the online and verbal discourse. Notably, students’ contributions on Knowledge Forum showed an apparent pattern of stronger connections among codes of higher semantic levels, suggesting that students were more cognitively engaged in the online discussion than their group verbal talk. Implications for research and practice are discussed.

Keywords: Knowledge Building, Discourse Analysis, Epistemic Network Analysis

1. Introduction

Knowledge Building (KB) is an established learning sciences theory that seeks to promote innovative ideas and idea improvement among students via collaborative engagement in productive discourse. Knowledge Building brings students together to work as a community to tackle real-world problems and creates a discursive environment that supports students to develop a deeper collective understanding of the problem at hand (Scardamalia & Bereiter, 2006). When a community of students is engaged in Knowledge Building discourse to advance collective understanding, they contribute important moves such as questioning, explaining with evidence, adding new information and so on, in order to advance the collective inquiry. The understanding of what and how students contribute to improve ideas and their engagement in KB discourse have been of interest among researchers. In fact, to understand student engagement in KB discourse, various frameworks have been proposed to characterise student discourse moves such as theory building moves (Chuy et al., 2010) and idea complexity (Zhang et al., 2007). Recent works suggest quantitative modelling techniques such as Epistemic Network Analysis (ENA) to assess connections between student moves (Ma et al., 2019). Insights from such analyses provide a means to understand the quality of student ideas from discussions as well as how they develop collective inquiry (build on ideas). However, current understanding on KB discourse remains limited to students’ online participation. Although small group discussion is a common practice, there is little understanding on the role of verbal discussions to support KB discourse. This paper attempts
to address this line of inquiry by assessing student engagement in KB discourse supported by both online and verbal discussions. Specifically, we aim to understand the similarities and differences in the students’ discourse moves and connections between their ideas as they participate in online and verbal discussions in KB lessons.

2. Literature

2.1 Knowledge Building discourse

Knowledge building discourse is characterised by the generation of novel and challenging student ideas, questions, and different perspectives that strive to improve the collective understanding of the problem at hand. It focuses on the generative and creative nature of a classroom discussion, moving away from a more typical and traditional classroom discussion that focuses on getting the single correct answer (Ong et al., 2021; Teo et al., 2022). The goal of Knowledge Building classrooms is beyond information sharing and divide and conquer to put a product together (Scardamalia & Bereiter, 2006). The essence of knowledge building classroom lies in the Knowledge Building discourse. In KB discourse, students are encouraged to constantly question and refine their ideas collectively, find their motivation to connect with the big problem and help each other progress in the inquiry and knowledge. KB discourse is usually supported by the Knowledge Forum which is an online platform to facilitate students in posting and building on ideas (Scardamalia & Bereiter, 2006). Figure 1 shows a view of the Knowledge forum where students interact and contribute to the discussion together. Each node (square box) represents a note (contains ideas) from a student, and the arrows show the build-on from one node to another. However, it is also common for students to work in small groups and discuss verbally to improve ideas. Figure 1 shows a group of four students in a face-to-face setting with computer devices to allow students to engage in both online and verbal discussions.

![Figure 1](image.png)

Figure 1. Face-to-Face and Online discussion modes in Knowledge Building

2.2 Understanding student engagement in KB discourse

Several ways of understanding student contributions in KB discourse have been proposed in the field. One notable area of focus is on the characterisation of students’ ideas or discourse moves (e.g., Chuy et al., 2010; Zhang et al., 2007; 2022). Notably, Chuy and colleagues proposed the theory building moves to code student discourse moves. Specifically, they developed an empirically grounded list of six major ways of contributing to a productive dialogue for KB discourse. These categories comprise: asking thought-provoking questions,
theorizing, experimenting, working with evidence, creating syntheses and analogies, and supporting discussion. From another study, Zhang and colleagues (2007) looked at elementary students’ (Grade 4) contributions in KB discourse from the quality of their individual ideas. They devised the idea complexity scheme to categorise students’ ideas. Questions may be categorised into explanation-seeking versus fact-seeking questions and single-area versus cross-area questions. According to Zhang et al. (2007), students’ contributions can also be assessed based on the level of scientific sophistication (pre-scientific; hybrid; basically scientific; and scientific) and the level of epistemic complexity (unelaborated facts; elaborated facts; unelaborated explanations; elaborated explanations).

Another area of focus is on the epistemic understanding via patterns of how students build on each other’s ideas to progress knowledge (e.g., Lin & Chan, 2018; Ong et al., 2021). For instance, Lin and Chan introduced the thread analysis to code the quality of student online discourse. The authors found and categorised different patterns of how students build on each other’s ideas into four epistemic levels of increasing collective knowledge advancement. The basic level, Fragmented Discussion, showed no connection or a weak connection between student ideas which do not contribute to knowledge advancement. The next level, Knowledge Sharing, showed connections between student ideas mainly for exchanging information and may not be essential to solving the problem. Knowledge Construction represented a higher level where there is a building of connections to generate questions and explanations to construct theories. Finally, Knowledge Building represents the highest epistemic level where there are coherent connections between ideas to bring about new understanding and progressive development of students’ theories. Lin and Chan’s work highlighted the importance and need to consider idea connection when assessing student engagement in KB discourse, however their coding scheme remains restricted by few epistemic levels.

Recent works suggest the potential of modelling techniques to understand connections between student discourse moves in Knowledge Building. For instance, Ma and colleagues (2019) explored the Epistemic Network Analysis (ENA) to assess Knowledge Building discourse, specifically looking into “ways of contributing” to the discourse such as questions and theories. The authors argued that there is potential for ENA to model contribution dynamics during cycles of knowledge advancement. Ma et al.’s study demonstrated the potential of ENA as a feasible technique to identify similarities and differences in discourse patterns between groups. For example, one group may differ from another in terms of certain connections such as theorizing and questioning moves. As explained by Ma and colleagues, such network models can be used to understand complementary engagement in group discussions and indications of collaborative engagement in both online and verbal discourse.

While there is an extensive development of techniques to assess student contributions in KB discourse, current understanding on KB discourse remains limited to students’ online participation. Existing studies tend to focus on student participation in online discussions supported by the Knowledge Forum. There is little understanding on the role of verbal discussions to support KB discourse, even though it is a common practice among teachers to plan for small group discussions to allow students to talk about their ideas verbally (see Ong et al., 2021). Elsewhere beyond KB contexts, there are studies showing that verbal discussions do not necessarily compliments nor supports online discourse. For example, a recent study by Bagheri and Zenouzagh (2021) highlighted that students engaged differently in face-to-face and online conversations. The researchers investigated the discourse from a group of 30 student participants and they found differences in verbal, paralinguistic and functional level between the students’ face-to-face and online conversations (computer-mediated communication). Their study revealed that student engagement in online discourse was more productive compared to the physical discourse as (i) the students showed active control of the talk and engaged in richer exchanges of ideas such as argumentation and (ii) students have more tendency to appeal to the teacher for help and displayed silences and pauses in the physical discourse (Bagheri & Zenouzagh, 2021). As student engagement in both online and verbal discussion may impact on the quality of their collective inquiry in KB, it is thus important for ongoing research to delve deeper on this understanding.

In this paper, we hope to build on this line of inquiry by assessing students’ ways of contributing in both online and verbal discussion to better understand their engagement during
Knowledge Building discourse. We report on the interactional patterns in verbal and online discussions from a group of students engaged in KB collective inquiry over a 2.5hr lesson. Specifically, we aim to understand the similarities and differences in the students’ discourse moves and connections between their ideas as they participate in online and verbal discussions in KB lessons by using Epistemic Network Analysis. Our research question therefore is “What connections of student contribution can we see during verbal and online discussion using ENA?”

3. Research design and Analysis

3.1 Methods

To capture students’ verbal interaction during face-to-face discourse, we used a 360-degree camera placed at the center of the group to video-record the group interaction as shown in Figure 4. We flattened the video data to enable us to see the interaction of each participant in the group. To ensure high quality of verbal data is captured, we used another audio recorder with lapel microphones for every individual student to capture their verbal data (Figure 2). This setup was part a multimodal recording setup that also collected students’ self-reports on their emotions and physiological data from Empatica E4 wearables. However, we focused on the description of the setup on the video and audio recording.

3.2 Lesson design and implementation

In this case study, we collaborated with a teacher (experienced in knowledge building) to implement a lesson designed with Knowledge Building activities to support collective inquiry and idea improvement. Table 1 shows the teacher’s planned lesson activities that engaged students to conduct idea generation and idea building about a real-world problem (based on Social Studies). Students will be tasked to play the roles of locals situated in three Southeast Asian areas (Sarawak Rainforest, Mekong River, and Anak Krakatoa) and they will continue to discuss the pros and cons of relocation and to provide their final stand with justifications.

Before the lesson, the students were pre-assigned to one of the three geographical regions. Each group was tasked to discuss and find solutions to existing problems (habitat loss, active volcano, and deforestation) that threaten the respective residents’ livelihoods and environments. Table 1 shows the flow of the activities that were planned to provide
opportunities for student discussion both online and face-to-face. A total of 17 Primary Five (Grade Five) students participated in this lesson that lasted for 2.5 hours.

Table 1. Activities from the teacher’s lesson plan

<table>
<thead>
<tr>
<th>Lesson phase</th>
<th>Teacher’s instructions and class activities</th>
<th>Students’ KF activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activating</td>
<td>● Students played the roles of locals in three Southeast Asian areas (Sarawak Rainforest, Mekong River, and Anak Krakatoa). ● Students introduced to the KB task, which is a Town Hall discussion anchored on the question: “Should our community stay or relocate from our area?”</td>
<td>No KF activities recorded</td>
</tr>
<tr>
<td>Idea Generation</td>
<td>● Students discussed the anchor question in their respective groups. ● Students posted notes on KF stating their stand and with supporting reasons.</td>
<td>Discuss and post different perspective (stand) with reasons on KF</td>
</tr>
<tr>
<td>Idea Improvement – Build On</td>
<td>● Students discussed and built on their peers’ ideas on the Knowledge Forum using build-on scaffolds. ● The teacher used learning analytics such as word clouds and an in-built Scaffold Tracker to conduct a meta-discussion and highlight value-adding contributions.</td>
<td>Read and build on groupmates’ viewpoints, such as clarifying, inquiring, or proposing new elaborations.</td>
</tr>
<tr>
<td>Idea Improvement – Rise Above</td>
<td>● Teacher introduced a Rise-Above scenario: Students took on newer roles as community representatives to potentially implement a solution to the respective area’s issues.</td>
<td>Negotiate as a group, weighing pros and cons of different solutions to develop a solution. Post synthesis of group ideas and peer feedback for improvement.</td>
</tr>
<tr>
<td>Idea Assessment</td>
<td>● Students consolidated their problems and solutions that were then shared with the class for comments and suggestions.</td>
<td>Post reflection on KF</td>
</tr>
<tr>
<td>Reflection</td>
<td>● Students responded to three reflection questions on KF.</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Analysis

In this study, we used the ENA Web Tool (Shaffer et al., 2009; Shaffer et al., 2016) to analyse both Knowledge Forum data and verbal discussion data. ENA models the weighted structure of connections in discursive data (Shaffer & Ruis, 2017). The method was designed to make sense of connections between ideas and actions using a repertoire of network-based methods that include visualizations and statistical modelling. For example, if a student group is discussing a particular topic, they share important elements such as the areas of their interest, their questions, information that they know, and so on. They may have a few conversations during the lesson, and a key part of understanding their discourse is modelling how they think about the connections between the different elements of their inquiry. ENA quantifies the co-occurrence of, for example, codes in discourse or elements of interaction in a chat and produces a weighted network visualization that shows connections for each unit of analysis (e.g., individual speakers or sub-groups of speakers) (Shaffer & Ruis, 2017). This paper reports the preliminary analysis with one student group (six students).
Based on 42 Knowledge Forum notes from the group, we coded for student contribution based on five levels of semantic patterns (adapted Zhang et al., 2022). These included casual talk (CT), fact-seeking question (Q1), explanation-seeking question (Q2), explanation with simple points or simple statement without elaboration (E1), and explanation with elaborations or rise-aboves (E2). Table 2 provides the description and sample coding. We use each student idea as the unit of analysis. This is typically in the form of a question or an explanation (statements). Where a student note or utterance contains a question and explanation, we coded for both respectively. We coded student contribution from their verbal talk as well. In total, we captured approximately 1.5 hr of talk time from the student group and this data was transcribed (2416 utterances) and coded with the same coding scheme and unit of analysis. In total, we coded 51 codes from the KF notes and 2548 codes from the verbal data.

We next applied Epistemic Network Analysis (ENA) to identify and quantify connections between these codes for the group. We created the epistemic networks by firstly using the discourse type (KF and verbal) as units for ENA to construct networks. The two networks were aggregated using a binary summation to reflect the presence or absence of the co-occurrence of each pair of codes within a stanza window of 6 lines in the conversations. The conversations are collections of lines that ENA uses to model connections between concepts, and this was based on the student utterances or KF notes in our analysis. Finally, codes are considered concepts and its patterns of association are what we want to model. In our case, we applied all the five semantic levels of the student contribution.

Table 2. Description and examples of semantic patterns of KF notes

<table>
<thead>
<tr>
<th>Codes of semantic patterns</th>
<th>Description/Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual Talk (CT)</td>
<td>Information or inputs generally irrelevant to the collective inquiry, e.g., “I need the toilet.”</td>
</tr>
<tr>
<td>Fact-seeking question (Q1)</td>
<td>Questions that elicit factual information, using when, where, or who, e.g., “Where do you want to go?”</td>
</tr>
<tr>
<td>Explanation-seeking question (Q2)</td>
<td>Questions which inquire into relations between facts and elicit elaborations, such as why, how, and what, e.g., “Why should we be affected by other people living in the modern world?”</td>
</tr>
<tr>
<td>Explanation with simple points/Simple statement without elaboration (E1)</td>
<td>Simple statement without elaboration, e.g., “You are contradicting yourself. These are the exact reasons why we should move.”</td>
</tr>
<tr>
<td>Explanation with elaborations/rise-aboves (E2)</td>
<td>Statement with elaborations to provide reasons, relationships, or synthesis of ideas, e.g., “We don’t really have a government. Without a government, there is no law, no order. And with a government, we suffer under a dictatorship. People who are weaker will abuse power.”</td>
</tr>
</tbody>
</table>

4. Findings

Answering the research question “What connections of student contribution can we see during face-to-face and online discussion using ENA?”, Figure 3 shows the epistemic network between the group’s verbal talk and KF discussion throughout the entire lesson. The positions of the network graph nodes are fixed, and the nodes correspond to the codes applied in the analysis. ENA determined these positions using an optimization routine that minimises the difference between the plotted points and their corresponding network centroids (Shaffer et al., 2016). The lines reflect the relative frequency of co-occurrence, or connection, between
two codes. The weight of each code is represented by the size of the node in the network graph. The results show that the group demonstrated a relatively unique profile between the two discourse types. Student contributions from their KF discussion shows centrality on elaborated questions (node with highest weight) and strongest connections between this code with explanations (simple and elaborated). Student contributions from their verbal discussion however shows centrality on simple explanation and strongest connections with casual talk (thickest line) and questions (simple and elaborated) than elaborated explanations. This preliminary analysis from this group suggest that students seem to engage in a deeper discussion of their ideas when they interact on KF compared to their verbal interaction (despite that they were given time to discuss verbally).

Figure 3. Epistemic networks showing connection of codes (semantic levels) for (a) verbal and (b) online discussion from the student group.

This finding was corroborated when we looked at the distribution of codes for the verbal and online discussion for the group. As shown in Figure 4, there was also a considerable difference in the distribution of codes between the two discourse modes. Student contributions in their KF discussion showed highest occurrence of simple explanation followed by elaborated explanations and then elaborated questions and lower occurrences of casual talk and simple questions. Student contributions in verbal discussion also showed highest occurrences of simple explanations but followed by casual talk and lower occurrences of the codes of higher semantic levels (elaborated explanations and elaborated questions).

Figure 4. Comparison of codes (semantic levels) between verbal and online discussion from the student group.

However, we found interesting observations when we further explore the epistemic network of individual students. As shown in Figure 5, 3 students in the group showed connections between the codes for their online discussion for the group but their network profile were not totally consistent. Specifically, two students, S1 and S5, showed strongest
connections between elaborated question and elaborated explanation (thickest line), while student S4 showed strongest connections between elaborated question and simple explanation and question. This result suggests that S1 and S5 were consistently using higher semantic levels like elaborated questions and explanations to deepen the collective inquiry. Note that the ENA did not picked up connections for the other students (as these students contributed very few notes).

Similarly, their individual network profile was also different in the verbal discussion. Figure 6 shows connections between the codes for their verbal discussion. This time, students S1 and S4 showed strongest connections between elaborated question and simple explanation, whereas students S1 and S5 displayed strongest connections between elaborated question and elaborated explanation (thicker line). The other three students S2, S3, and S6 showed weaker connections between these codes but strongest connection between simple explanation and casual talk. Thus, students S1, S4 and S5 seemed more engaged in a deeper discussion of their ideas during their verbal discussion, which is a consistent find with their contributions on KF discussion.

5. Discussion and conclusion

Our preliminary finding from the ENA analysis is consistent with those by Ma et al. (2019) as we found that students deepened their collective inquiry as they engaged in Knowledge Building discourse. As shown in our analysis, this group of students reflected a strong connection between higher semantic levels of contributions elaborated questions and elaborated explanations when given time to discuss on KF. However, we also found different patterns of their contributions during verbal discussion with the ENA analysis. Specifically, in their group verbal talk, there was a strong presence of casual talk and connections between
simple explanations and questions which suggest that the collective inquiry was more superficial during verbal discussions. This claim that students engage more productively in online communication is consistent with the report from Bagheri and Zenouzagh (2021). However, we think that such a result may also be due to the affordances of the online KF platform such as the support of KB scaffolds and the chance to read (even multiple times) and consider multiple peer ideas. Unlike KF, students tend to observe turn-taking during verbal discussion, which may restrict them in responding to certain ideas and giving longer responses such as elaborated explanations. Therefore, future work can consider looking into the impact of different support features in both online and verbal discussion.

An interesting finding was the interactional patterns from individual students. The ENA analyses suggested that students who were engaged in a deeper discussion of their ideas on KF also contributed ideas of higher semantic levels during their verbal interaction. This finding has an implication for practitioners particularly on the lesson design and facilitation. For example, teachers can first plan for an online discussion to allow students to generate questions or explanations of higher semantic levels and then allowing them to continue building their ideas via small group discussions. As it is easy to extract student notes from the KF, teachers can also run the data using the ENA tool and use the analysis to discuss with students on how their ideas are connected and ways to improve the quality of their talk. Such teacher facilitation has been shown to be beneficial for students to develop new inquiry (Zhu & Lin, 2023). However, we recognise that our findings were based on a very small dataset and our claims warrant further investigation with more data.

Moving forward, we have collected new datasets, and additional analyses are underway to compare findings from other groups and to validate the claims. In conclusion, this study highlights that the understanding of student engagement in KB discourse is not straightforward, as student interactions can occur within and between groups as well as via online or verbal discussion. As we found that the type of and connections between student discourse moves were less sophisticated in their verbal discussions, a key implication from this study is that students need more support during verbal talk to engage in deeper collective inquiry and advancing the community knowledge. Our study also shows ENA as a promising approach to assess Knowledge Building discourse in both online and verbal mode. ENA affords modelling of both group and individual networks to enable the understanding of differences in student engagement during KB discourse. Future studies can explore on the automation of such feedback for teachers and students. For example, the real-time capturing and coding of verbal data which is integrated with the ENA tool to generate feedback on the discourse quality from the verbal discussion.

Acknowledgements

This study was funded by the Ministry of Education (MOE), Singapore, under the Education Research Funding Programme (OER 19/19 TCL) and administered by National Institute of Education (NIE), Nanyang Technological University, Singapore. The views expressed in this study are the authors’ and do not necessarily represent the views of the host institution. The authors thank the teacher and student participants involved in this study. The authors also express their appreciation to Associate Professor Justin Dauwels and his team for expert advice and support on the multimodal recording setup.

References


Investigating Trustworthiness and Conflict in Historical Multiple Texts: From Eye-Tracking Data of Source and Content Processing

Zheng-Hong Guan\textsuperscript{a}, Sunny S. J. Lin\textsuperscript{b}  
\textsuperscript{ab}National Yang Ming Chiao Tung University, Taiwan  
*sunnylin.nctu@gmail.com

Abstract: Online reading is one of the sources to acquire knowledge, but learners may encounter conflicting texts during reading. Thus, learners need to distinguish the source's trustworthiness and determine which perspectives to believe. However, previous studies have focused more on the source than on the content and more on science than on history in eye-tracking fields. Therefore, the purpose of this study is to investigate how readers perceive conflict and process source and content when reading historical multiple texts with varying degrees of trustworthiness. 11 participants were recruited and presented six historical texts under three different conditions: high-trustworthiness, low-trustworthiness, and trustworthiness-differences. Their eye movement was recorded. The results indicated that the readers rated the low-trustworthiness condition as having more conflict than the others. Additionally, the low-trustworthiness condition had less first-pass reading time than the trustworthiness-differences condition. Readers spent more rereading time on the content than on the source only in the trustworthiness-differences and low-trustworthiness conditions. When reading text with low trustworthiness, the reader spent more first-pass reading time on the source than on the content and spent more rereading time on the content than on the source. These findings have several implications. Firstly, readers were overwhelmed and needed guidance on which information to believe, leading to the perception of more conflict when reading low-trustworthiness conditions. Second, the readers were aware that the source seemed untrustworthy, leading them to decrease reading time to process the following content. After confirming the source's trustworthiness, readers tended to reread the content to resolve the conflict. Lastly, low-trustworthiness sources decreased processing time in the following content, indicating that readers regulate their reading patterns depending on the text's trustworthiness. Overall, this study provides insights into how readers process historical multiple texts with varying degrees of trustworthiness.

Keywords: multiple-text, eye-tracking, conflicting, historical topic

1. Introduction

As technology advances, people can get much but conflicting information on the Internet. To comprehend an issue deeply, readers need to read more than one text to validate the trustworthiness of the information (Goldman & Scardamalia, 2013; Jarodzka & Brand-Gruwel, 2017). When reading multiple texts, readers need to select, process, and evaluate the relevant ones to form meaningful mental representation (Braten & Strømsø, 2009).

In the context of multiple-text use, reading strategies are associated with evaluating the trustworthiness (Braasch & Braten, 2017). When reading conflicting information, readers perform better if they evaluate the source’s authorship, intent, and related affiliation (Anmarkrud et al., 2014; List & Alexander, 2019). However, content processing is equally essential but seems to be neglected in the eye-tracking field (Salmerón et al., 2018). When judging trustworthiness, readers should rely on both the source and content simultaneously.
(Barzilai et al., 2020). Therefore, the aim of this study is to extend previous research by investigating the effect of trustworthiness differences, content, and source on eye movement processes when reading conflicting information.

2. Literature Review

2.1 Conflicting information in multiple texts

The information between texts can be a complementary, redundant, or conflicting relationship, and the conflicting information has been the focus of much research attention (Primor & Katzir, 2018). Previous research has proved that readers pay attention to the source information and memorize them (Braasch et al., 2012; Kammerer et al., 2016; Rouet et al., 2020). Braasch and Bråten (2017) proposed the discrepancy-induced source model (D-ISC model) to explain this phenomenon. The incoherence provoked by conflicting information leads readers to focus on source information to explain why the conflict happened, resulting in deep memory encoding of the source information.

When encountering conflicting information, evaluating the source’s trustworthiness is one of the approaches to reconciling this situation. For instance, Bråten et al. (2011) instructed readers to read various texts on climate change from different sources, including textbooks, newspapers, magazines, and official documents. The readers rated textbooks and official documents as trustworthy and paid more attention to these sources than others. Gottschling et al. (2019) required the experimental group to read two conflicting scientific texts, one written by a professor working for a company and the other by a university professor, while a control group read two scientific texts written by university professors. Their results revealed that readers rated the text from the company as less trustworthy than the text from the university. Previous research on this topic seems to focus more on scientific than historical issues which need further investigation.

2.2 Eye-tracking in multiple-text reading

Eye-tracking has been used to investigate the moment-to-moment attentional distribution (Rayner, 2009). Researchers usually define the source information as the area of interest to examine readers’ attention when reading the conflicting information (Braasch et al., 2012; Gottschling et al., 2019; Kammerer et al., 2016). For example, Kammerer et al. (2016) explored whether readers paid attention to processing the source information when reading conflicting information on web pages. The result revealed that readers had longer total fixation duration on the source information when reading conflicting web pages than consistent ones. Gottschling et al. (2019) used first-pass reading time and rereading time to examine readers’ cognitive process when reading conflicting information. Their findings revealed that the trustworthiness differences in information caused readers to spend more time rereading the low-trustworthiness source.

These studies suggest that readers allocate attentional resources to source information when processing conflicting information. However, the discussion about content processing seems less in the eye-tracking field of multiple-text use.

2.3 The current study

The conflicting multiple texts make readers concentrate on the source information for resolving the conflict. Previous studies have been conducted with scientific texts. It needs to be clarified how readers process multiple historical texts, especially considering the content and sources simultaneously. Therefore, the aim of this study is to investigate readers’ attentional distribution when reading multiple historical texts with conflicting information. The trustworthiness of the sources was manipulated into three conditions: high-trustworthiness, low-trustworthiness, and trustworthiness-differences conditions. Two research questions were proposed:
RQ1: How conflict do readers perceive when reading different multiple-text conditions?
RQ2: How do readers process the source and content reflected from the eye-movement data when reading different multiple-text conditions?

3. Method

3.1 Participants

It was an exploratory study, and only 11 participants were recruited from the University in Northern Taiwan. All the participants passed the calibrations. This study was conducted with a within-factorial subject design. Every participant was instructed to read all conditions. The present order was counterbalanced across the participants.

3.2 Materials

The participants were given six texts regarding the same historical topic. Table 1 presented detailed information about these texts: either high trustworthiness, meaning that the source was from the government, or low trustworthiness, meaning that an anonymous person paraphrased the source. Three texts shared the same viewpoints, while the other three opposed those viewpoints. Each text consisted of two paragraphs: source and content. The source was positioned before the content, as presented in Figure 1. The six texts were paired into three conditions: high-trustworthiness, low-trustworthiness, and trustworthiness-differences. Two texts were displayed on the screen at the same time.

<table>
<thead>
<tr>
<th>Table 1. Overview of the six texts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>No. 1</td>
</tr>
<tr>
<td>Viewpoints</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Number of words in the</td>
</tr>
<tr>
<td>source/content</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Figure 1. The display of text condition.

3.3 Apparatus, area of interest, and eye movement measures

The eye movements were recorded by an EyeLink 1000 desktop remote eye-tracker system (SR Research Ltd., Canada) with a sample rate of 1000Hz. The distance between the monitor and the participants was 60 cm. Each Chinese character was 24 × 24 pixels, and the space between lines was 32 pixels. Figure 1 presented four AOIs (areas of interest) defined in each trial: two sources and two content. Two eye-movement measures were first-pass
reading time (FPRT) and rereading time (RRT). FPRT is the sum of the durations of all first-pass fixations before leaving the AOI. RRT is the sum of the durations of all returning fixations after first-pass fixations on the AOI. To control the length of paragraphs, eye-movement measures were divided by the number of characters (Stadtler et al., 2020).

3.4 Measures

After reading each condition, the participants were instructed to answer three questions using a 5-point Likert scale. One of the questions inquired about the sense of conflict between the two texts, while the other two questions referred to judging the trustworthiness of each text.

3.5 Procedure

The procedure in the experiment consisted of two sections. In the first section, the participants were provided with concise instructions on the eye-tracking procedure. They were positioned in front of a monitor with a chin rest to avoid head movements. The participants were calibrated with nine points. In the second section, participants were instructed to read three conditions. In each condition, the texts were conflicting with each other. Once the participants finished each condition, they were required to rate each text’s sense of conflict and trustworthiness. After the completion of the tasks, the participants were dismissed.

4. Result

4.1 The rating of trustworthiness and sense of conflict

To answer RQ1, we used repeated measure t-test to examine the trustworthiness in each condition and repeated measures ANOVA to examine the differences between the multiple-text conditions in the sense of conflict. The means and standard deviations (in parentheses) are presented in Table 2.

The result of the t-test showed a significant difference in the trustworthiness-differences condition, indicating the second text had high trustworthiness than the first text ($t(10) = -3.31$, $p < .01$). Additionally, the high-trustworthiness condition had higher trustworthiness than the low-trustworthiness condition ($t(10) = 3.46$, $p < .01$).

The results of RM-ANOVA showed that the sense of conflict differed significantly across the multiple-text conditions ($F(2,20) = 4.93$, $p = .018$, $\eta^2 = 0.16$). The post-hoc comparisons showed that the low-trustworthiness condition made readers perceive more conflict than the high-trustworthiness condition ($t = 3.12$, $p < .05$), while other comparisons were not significant.

<table>
<thead>
<tr>
<th>No.</th>
<th>Trustworthiness-differences condition</th>
<th>Low-trustworthiness condition</th>
<th>High-trustworthiness condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.27 (0.79)</td>
<td>3.18 (1.08)</td>
<td>3.00 (0.89)</td>
</tr>
<tr>
<td>3</td>
<td>3.91 (0.94)</td>
<td>2.45 (0.69)</td>
<td>3.73 (1.01)</td>
</tr>
</tbody>
</table>

4.2 Eye-movement measures

To answer RQ2, we analyzed two eye-movement measures using linear mixed model (LMM) with lme4 package (Bates et al., 2015). In this study, participants and texts were regarded as random effects. Two fixed effects were the multiple-text conditions with three levels (trustworthiness-differences, low-trustworthiness, and high-trustworthiness) and the
paragraphs with two levels (source and content). Two dependent variables were first-pass reading time and rereading time. Table 3 presents the means and standard deviations (in parentheses) for eye-movement measures in each condition.

Table 3. Mean and SD for eye-movement measures in each condition of texts

<table>
<thead>
<tr>
<th>Multiple-text conditions</th>
<th>Trustworthiness-differences condition</th>
<th>Low-trustworthiness condition</th>
<th>High-trustworthiness condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Source</td>
<td>Content</td>
<td>Source</td>
</tr>
<tr>
<td>First-pass reading time</td>
<td>41.24</td>
<td>29.67</td>
<td>26.46</td>
</tr>
<tr>
<td></td>
<td>(46.81)</td>
<td>(39.47)</td>
<td>(32.27)</td>
</tr>
<tr>
<td>Rereading time</td>
<td>60.22</td>
<td>100.40</td>
<td>66.19</td>
</tr>
<tr>
<td></td>
<td>(55.87)</td>
<td>(55.01)</td>
<td>(64.32)</td>
</tr>
</tbody>
</table>

The result of first-pass reading time showed that the low-trustworthiness condition had less first-pass reading time than the trustworthiness-differences condition ($b = -14.98$, $SE = 6.9$, $t = -2.16$, $p < .01$). Other comparisons were not significant.

The result of rereading time showed that the main effect of the paragraphs was significant ($b = 27.28$, $SE = 8.22$, $t = 3.31$, $p < .01$), indicating that the readers spent more time on the content than on the source. However, this effect was moderated by the multiple-text conditions because of the two-way interaction effect ($b = -43.40$, $SE = 8.22$, $t = -2.15$, $p < .05$).

![Figure 2. The interaction effect of multiple-text conditions and paragraphs on rereading time.](image)

In Figure 2., post-hoc comparisons showed that readers spent more time on the content than on the source only in the trustworthiness-differences condition ($b = -40.18$, $SE = 14.2$, $t = -2.82$, $p < .01$) and the low-trustworthiness condition ($b = -44.50$, $SE = 14.2$, $t = -3.12$, $p < .01$). These findings were inconsistent with previous studies, which revealed that when reading two conflicting texts, readers would shift their attention to the source to reconcile the conflict (Braasch et al., 2012; Kammerer et al., 2016). We speculated that readers would regulate their depth of reading patterns based on the source's trustworthiness. If readers were aware of the untrustworthy texts, readers might decrease the processing time in the following content. To test this hypothesis, we merged these three conditions to analyze these six texts with either high or low trustworthiness, and each text contained two paragraphs. In other words, two independent variables (trustworthiness and paragraphs) and two dependent variables (first-pass reading time and rereading time) were analyzed.

In Figure 3., the result of first-pass reading time (left side) showed the marginally significant interaction effect ($b = 21.30$, $SE = 11.17$, $t = 1.90$, $p = .06$). The post-hoc comparisons revealed that when reading the texts with low trustworthiness, readers spent more time on the source than on the content ($b = 18.59$, $SE = 7.9$, $t = 2.35$, $p < .05$) while this
effect was not found in reading the text with high trustworthiness ($b = -2.71$, $SE = 7.9$, $t = -0.34$, $p = .73$). The result of rereading time (right side) showed the significant main effect of paragraphs, indicating that readers spent more time on the content than on the source ($b = 27.15$, $SE = 8.30$, $t = 3.27$, $p < .01$). The interaction effect also reached a marginally significant level ($b = -31.40$, $SE = 16.60$, $t = -1.89$, $p = .06$). The post-hoc comparisons showed that when reading the texts with low trustworthiness, readers spent more time on the content than on the source ($b = 42.90$, $SE = 11.70$, $t = -3.65$, $p < .001$) while this effect was not found in reading the texts with high trustworthiness ($b = -11.4$, $SE = 11.70$, $t = -0.97$, $p = .33$).

![Figure 3](image)

**Figure 3.** The interaction effect of texts trustworthiness and paragraphs on first-pass reading time (left side) and rereading time (right side).

5. **Conclusion**

The previous studies have found more on the source than on the content (Gottschling & Kammerer, 2021; Gottschling et al., 2019); however, the content is also essential for forming an elaborative document model (Perfetti et al., 1999; Salmerón et al., 2018). Therefore, the goal of this study is to investigate readers' attentional distribution when reading a historical topic with conflicting pieces of information varied with trustworthiness.

Three main findings were drawn from the analysis. Firstly, readers rated the low-trustworthiness condition as having more conflict than the other conditions, suggesting that they were overwhelmed and unsure which information to believe; hence, they perceived more conflict. The other conditions provided readers with at least one text to believe, which lessened readers' sense of conflict. Future research should investigate how readers judge conflict.

Second, the eye-movement analysis indicated that the low-trustworthiness condition had less first-pass reading time than the trustworthiness-differences condition, implying that readers were aware that the source seemed less trustworthy; therefore, they decreased reading time to process the following content. This trend was also evident in the mean of first-pass reading time in Table 3, although the difference was not significant. Additionally, the result of rereading time revealed that readers spent more time on the content than on the sources, particularly when reading the trustworthiness-difference and the low-trustworthiness conditions. This suggests that readers tend to reread the content after confirming the source's trustworthiness to resolve the conflict.

Thirdly, we merged these three conditions into six texts with either high or low trustworthiness. We found that only low-trustworthiness texts caused readers to increase first-pass reading time processing the source and rereading time processing the content. This effect was not found in reading high-trustworthiness texts. In other words, low-trustworthiness sources led to decrease processing time in the following content, indicating that readers regulate their reading patterns depending on the text's trustworthiness. This explanation echoed the second finding and should be confirmed by future research.

Some limitations should be noted. Firstly, only 11 participants were recruited. Accordingly, some statistical results only reached marginal significance. Therefore, the sample should be increased to confirm the statistical results. Secondly, we did not collect reading comprehension or memory data, which is essential in the document model.
emphasizing the link between the source and the content (Perfetti et al., 1999). Future research should evaluate readers’ reading comprehension. In summary, this study provides empirical eye movement data and insight into how readers process sources and content in a historical topic. Future research should continue to explore these issues.

References

Anmarkrud, Ø., Bråten, I., & Strembsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, 30, 64-76. [https://doi.org/10.1016/j.lindif.2013.01.007](https://doi.org/10.1016/j.lindif.2013.01.007)


Bråten, I., & Strembsø, H. I. (2009). When law students read multiple documents about global warming: examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instructional Science*, 38(6), 635-657. [https://doi.org/10.1007/s11251-008-9091-4](https://doi.org/10.1007/s11251-008-9091-4)


Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Q J Exp Psychol (Hove)*, 62(8), 1457-1506. [https://doi.org/10.1080/17470210902816461](https://doi.org/10.1080/17470210902816461)


Multimodal analysis of learners’ communications in CSCL of a mathematical proof

Masataka KANEKOa*, Hironori EGIb & Takeo NODAc

aFaculty of Pharmaceutical Sciences, Toho University, Japan
bGraduate School of Informatics and Engineering, The University of Electro-Communications, Japan
cFaculty of Science, Toho University, Japan
*masataka.kaneko@phar.toho-u.ac.jp

Abstract: In this study, we explore the CSCL of a mathematical proof using HTML-based dynamic geometry content. In addition to the learners’ transcribed communication, multimodal log data of their manipulations of the content and utterances are analyzed. First, we use multimodal data to segment the whole learning process into stages and identify the characteristics of those stages. This segmentation is then aligned with the transcribed communication. Our finding is that the participants’ choice of words varies according to each stage and is compatible with the characteristics derived from multimodal data. Resultingly, when and how the convergent conceptual change occurred is inferred from the commognitive perspective.

Keywords: Mathematical proof, dynamic geometry, log data of manipulations, log data of utterances, text mining of annotated dialogue

1. Introduction

While mathematicians construct proofs as part of their professional work, students also construct them in mathematics classes. In both situations, some scientific discovery is involved and the convergent conceptual change leading to it is often investigated through ethnomethodological studies (Garfinkel, 2002). One puzzle related to the investigation of scientific discovery lies in how is it that we can talk about the matter being discovered while still engaged in the work of discovering it (Koschmann, & Zemel, 2009). Moreover, especially in the case of the activities in classroom, the account of the discovery tends to become highly complex due to the learners’ lack of domain-specific knowledge and related vocabulary. In their previous work (Kaneko, Egi, & Noda, 2022), the authors investigated learners’ collaborative activity of proving the addition theorem for trigonometric function while manipulating HTML-based content generated using the dynamic geometry system CindyJS (https://cindyjs.org/). For that investigation, a Moodle plug-in of CindyJS was used to store the log data of learners’ manipulations (Kaneko, Nakahara, & Noda, 2020). Moreover, the conversation analysis system Diana-AD was also used to record learners’ utterances and measure the loudness of their voices (Ishikawa, Okazawa, & Egi, 2019). One of the main findings was that high-resolution multimodal log data derived from them could serve as well-grounded material for inferring where to “locate a proposal for a possible discovery” in the whole learning process (Koschmann, & Zemel, 2009). Based on that finding, in this study, we explore another group’s learning in a similar setting and discuss how to align the multimodal log data with the transcribed communication. In the case of this proof activity, learners can be assumed to experience some commognitive conflict while generalizing their prior knowledge to newly emerging situations. Their participation in some discursive processes (metadiscourse) to handle the relevant commognitive conflict can lead to the reification of those processes through which increasingly sophisticated approximations to the targeted concept are constructed (Sfard, 2009). Since this reification can be assumed to
cause some change in the pattern of both manipulation and discourse, we first segment the whole learning process according to the change in their manipulating pattern and the power balance between each participant’s utterances. Then, we examine the difference in each participant’s choices of words and their associated meanings among the segmented parts.

2. Research Methods and Results

Figure 1 (left) shows the content used in which the angles $\alpha$ (red) and $\beta$ (green) can be moved. Two first-grade female students with a pharmaceutical science major in a Japanese university were asked to manipulate the content collaboratively on an iPad. Beforehand, they were asked to fill in the worksheet in Figure 1 (center and right) to prove the theorem for the rudimentary case where $\alpha + \beta$ was acute. The time they needed to do that was measured and one participant (A) was assumed to have higher prior skill than the other (B).

![Figure 1. The CindyJS content (left) and the worksheet (right) including the figure (center)](image)

While Figure 2 (left) shows the fluctuations of the three angles $\alpha$ (blue), $\beta$ (orange), and $\alpha + \beta$ (grey), the yellow curve in Figure 2 (right) shows the power balance of utterance between A (lower domain) and B (upper domain). It can be seen from the left figure that, during 2:30-13:00, almost no manipulation was made. In fact, a videotaped image of the participants’ activities tells us that almost all of their motions were pointing to the figures in both worksheet and iPad during this time interval. Thus, we divide the whole process into three stages, Stage I (0min00sec - 2min30sec), Stage II (2min30sec – 13min00sec), and Stage III (13min00sec – 17min00sec).

![Figure 2. The fluctuation of angles (left) and the power balance of utterance (right)](image)

Aligning these results with the transcribed communication, it can be seen that A played the leading role in the search for a probable positional relationship between angles while manipulating the content at Stage I. Then, at Stage II, B gradually increased her presence while comparing the situation they encountered with the rudimentary case on the worksheet. At this stage, A became dominant only once between 6min00sec and 8min00sec when A presented some assumptions. Finally, at Stage III, A again became dominant when she claimed the orthogonality of $\perp$BHO, which is one of the key points.

To examine the participants’ choice of words, the author used the text mining software KHCoder (https://khcoder.net) which is specialized for the Japanese language. Some previous studies pointed out that a seemingly vague pronoun, which does not refer to any prior antecedent but is a prospective indexical, might leave room for later elaboration leading to a new understanding (Koschmann, & Zemel, 2009). Therefore, we calculated the frequency of the use of demonstrative pronouns (DP) and common nouns (CN) for each participant and stage. Figure 3 visualizes the resulting cross table in which the sizes of the
rectangles represent the incidence of sentences including DP (upper row) and CN (bottom row) among all ones in each category. As is seen, while B used DP throughout the whole process, A gradually replaced the use of DP with the use of CN.

![Figure 3. Visualization of each participant’s choice of words](image)

3. Discussion and Future Works

While most of the DP referred to some specific antecedents, there do exist some DP without any clear referent. For instance, after B said cautiously at 14min44sec, “The length of OH is, \( \cos \beta \)? I am not sure, but it will be the same,” A responded at 15min03sec, “Anyway, our task should be to check whether that still holds or not when the angles are changed.” It can be judged from the context that the pronoun “that” in this utterance of A was used not to refer to some specific fact but to propose the possibility of literally imitating the descriptions in the rudimentary case. Thus, this pronoun use clearly indicates that the preceding discourse had been reified into some newly proposed “discovery”.

Because of the limited number of cases examined, the findings in this study cannot be generalized to other cases directly. In fact, in another case study where the prior skill of the participants was fairly higher than that in this study, the utterances and manipulations of one participant were almost always dominant and the other participant only listened to her partner, occasionally throwing in an appropriate word or two. Therefore, many more cases should be explored to search for the possible communication patterns in this proof activity. Moreover, some situations in which “actions speak louder than words” were observed. The fine-grained log data of learners’ gestures should be synchronized and analyzed.

Acknowledgements

This work was supported by JSPS KAKENHI 19K03175, 21K03175, and 22K02951.

References


Development and Evaluation of a Matching System to Facilitate Online Collaborative Learning

Haruka TADA\textsuperscript{a} & Fumihide TANAKA\textsuperscript{b}
\textsuperscript{a} Graduate School of Science and Technology, University of Tsukuba, Japan
\textsuperscript{b} Institute of Systems and Information Engineering, University of Tsukuba, Japan
\textsuperscript{*}haruka.tada@fsl.iit.tsukuba.ac.jp

Abstract: Collaborative learning in programming education is known to have educational benefits and is now practiced worldwide. However, the quality of collaborative learning varies depending on the extent to which teachers can support it. Currently, teachers are often too busy to provide sufficient support for collaborative learning. Therefore, in this study, we focused on one form of support, learner matching, and developed a matching system that uses emotion recognition results from learners’ natural speech. We conducted comparative experiments with existing methods to clarify whether group formation based on the results of heterogeneity in emotion recognition is more effective than the existing method. The results suggest that the proposed matching system has the potential to bring positive effects to those learners with lower foundational skills. In addition, in supplementary surveys, the majority of participants reported that they found it easier to learn with their matched partners using the proposed method, thus supporting the hypothesis that the proposed method is more effective than the existing method.

Keywords: Collaborative Learning, CSCL, STEM Education, Programming Education, e-Learning

1. Introduction

The importance of science, technology, engineering, and mathematics (STEM) education is increasing, and programming education is being promoted actively around the world (Sáez-López et al., 2016). Programming education has been shown to improve students’ cognitive skills, such as creativity, reasoning, and mathematical skills, and collaborative learning is recommended in such education (Scherer et al., 2019). Collaborative learning is a teaching method in which students at various levels cooperate in small groups toward a common goal. Since they are responsible not only for their own learning but also for each other’s learning, the success of one student can then help others succeed (Gokhale, 1995). Collaborative learning has educational benefits over individual learning, and students who learn through collaborative learning are known to be able to engage in learning sustainably and persistently (SAHIN et al., 2014).

However, collaborative learning is not always effective, and its effectiveness depends on the richness and intensity of the interactions between group members during the collaboration (Dillenbourg & Hong, 2008). The presence of a teacher is also necessary to shape the interactions within the group (Muirhead, 2000). However, given the current situation where educational disparities are widening worldwide, it can be difficult for teachers to provide sufficient support. In Japan, educational disparities have arisen due to teachers’ increased workload. In Japan, programming education became mandatory in elementary schools in 2020 to foster programming thinking skills. However, in a survey of elementary school teachers, most of the respondents felt overwhelmed due to overwork, and 67 out of 79 respondents felt anxious about teaching programming education (Oshima et al., 2020). Therefore, in such a situation, it is difficult to provide programming education equally.
In the field of computer-supported collaborative learning (CSCL), several approaches support group interaction with computers. One of these approaches is group formation (Dillenbourg & Hong, 2008). Although group formation is important in collaborative learning, it is often overlooked, but proposing an appropriate group formation may make it possible to create beneficial groups (Muehlenbrock, 2006). Recent studies have used survey or test results, such as age, ability, and gender differences as criteria for group formation (Topping, 2010; Uchida, 2016). However, this is not necessarily related to the interactions between the group members, and it is necessary to consider heterogeneity regarding learners’ characteristics, such as their personality and social attitudes (Kinjo, 2018).

Therefore, we assume that the heterogeneity of the results of emotion recognition analyzed from the natural utterances of learners is relevant for group interactions. In this study, we targeted the scenario of pair work in robot programming and developed a system of matching by the heterogeneity of emotion recognition results that automates group formation in collaborative learning. In addition, as a research question, we aimed to clarify whether group formation based on the heterogeneity of emotion recognition results is more effective than current conventional methods. The system was evaluated through experiments with college students, who are easily accessible for feedback.

2. Overview of a Matching System

We developed a matching system using emotion recognition results from natural speech data acquired from pre-lesson interactions as an indicator. To procure natural speech from learners, they will learn with a virtual partner in a pre-lesson as shown in Figure 1. The virtual partner is an avatar programmed in advance for speech that talks to the learner like a real child, saying things like “What should we do here?” or “Have you finished yet?” to match the mission of the pre-lesson. By recording conversations between the virtual partner and the learner, natural speech data from the learner’s speech can be collected.

The reason for using a virtual partner rather than an actual person as a pre-lesson partner is to prevent changes in the learner’s speech content in response to the characteristics of the partner. By using a virtual partner, speech data can be collected equally under the same conditions for everyone. These data are then analyzed, and a matching algorithm is used based on the analysis results to match learners with each other. The matching algorithm was created through the experiments in Chapter 3.

![Figure 1. Matching Learners Based on the Analysis of Data from the Pre-Lesson.](image)

3. Development of a Matching System

The experiment involved a total of 12 participants (6 males and 6 females) who were university undergraduate and graduate students. In all subsequent experiments in this study, individual names were not identifiable to by third parties, participation was voluntary, and there was no penalty for refusing to participate. Participants were informed of the purpose and content of this study and provided their verbal and written consent. This study was approved by the Ethics Committee of the University of Tsukuba (2022R676).
3.1 Participants Were Randomly Matched

We randomly matched 12 participants and had them take a lesson in robot programming. The learning material for robot programming used in the experiment involved the use of micro:bit (micro:bit, 2023). Micro:bit is a small computer created by the British public broadcaster BBC to teach programming to children. As this was an online experiment, participants programmed the micro:bit using a simulator.

3.2 The Pre-Lesson for Participants and Virtual Partner/Analysis of Speech Data

Each participant was paired with a virtual partner for learning, and their speech data were collected during this process. The data were analyzed through emotion recognition as shown in Figure 3. First, those parts of the speech data that exceeded a certain amplitude for more than 0.6 seconds were identified as speech, and these segments of audio were extracted. Emotion recognition was performed using these speech data by inputting it into an emotion prediction model. The emotion prediction model was created using deep learning and the Ryerson Audio–Visual Database of Emotional Speech and Song (Livingstone & Russo, 2018), which is a dataset of 1440 audio samples to represent eight emotional states (neutral, calm, happy, sad, angry, fearful, disgusted, and surprised) stored in the data repository, Zenodo. The model structure uses Long Short-Term Memory, and the Mel-frequency cepstral coefficients (MFCC) were used as the feature. The reason for using MFCC as the feature is that it is effective for emotion recognition (Sato & Obuchi, 2007). The model was trained using training data, and 32 rounds of machine learning were performed to create an emotion prediction model with an accuracy of about 60% on test data.

Using this emotion prediction model, emotion recognition was performed. When the speech data of participant $p$’s $i$-th speech was inputted into the emotion prediction model, an array $x_p$ such as equation (1) was outputted. Here, $a_p(i)$ represents the neutral emotional level, $b_p(i)$ represents the calm emotional level, $c_p(i)$ represents the happy emotional level, $d_p(i)$ represents the sad emotional level, $e_p(i)$ represents the angry emotional level, $f_p(i)$ represents the fearful emotional level, $g_p(i)$ represents the disgusted emotional level, and $h_p(i)$ represents the surprised emotional level. For example, if $x_p(i) = [0.9, 0, 0.02, 0.04, 0, 0.04, 0, 0]$, it can be understood that the level of neutral emotion was high in the inputted speech data. These arrays were then averaged to create a single array. If the number of extracted speech data segments from participant $p$ is $n_p$, then the averaged array $\bar{x}_p$ can be obtained as shown in equation (2). This $\bar{x}_p$ was used as the emotion recognition results for participant $p$. 
\[ x_p(i) = \left[ a_p(i), b_p(i), c_p(i), d_p(i), e_p(i), f_p(i), g_p(i), h_p(i) \right] \]  

\[ X_p(i) = \left[ \frac{1}{n_p} \sum_{i=1}^{n_p} a_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} b_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} c_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} d_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} e_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} f_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} g_p(i), \frac{1}{n_p} \sum_{i=1}^{n_p} h_p(i) \right] \]

\[ = [A_p, B_p, C_p, D_p, E_p, F_p, G_p, H_p] \]  

**Figure 3.** Analysis through Emotion Recognition Using Deep Learning.

### 3.3 Lesson Between Participants/Evaluation of Pair Work

Participants were paired up to engage in learning activities, and evaluations of their pair work were conducted. In assessing pair work, Uchida used activation and satisfaction as variables to verify the effectiveness of collaborative learning (Uchida, 2016). We use these two scales.

Activation was defined as the number of speech units, defined as a cohesive segment of speech, and counted from text data obtained by transcribing recorded conversations during pair work. Uchida used activation as an evaluation metric for cooperative learning, as it is believed that the number of utterances counted is effective in estimating the state of cooperative work and provides useful information for assessing judgments of pairs with poor cooperative relationships and interactions between group members. We divided the length of the class period into seconds because different pairs had different study times. The activation for participant \( a \) and participant \( b \) was denoted by \( y_{a,b} \).

In this experiment, pairs were matched as follows: (participant 0 and participant 1), (participant 2 and participant 3), (participant 4 and participant 5), (participant 6 and participant 7), (participant 8 and participant 9), and (participant 10 and participant 11), and the resulting data set of activation was \( \{y_{0,1}, y_{2,3}, y_{4,5}, y_{6,7}, y_{8,9}, y_{10,11}\} \).

<table>
<thead>
<tr>
<th>Satisfaction Questionnaire Evaluation Items (UCHIDA et al., 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

Satisfaction was measured by a post-work survey that totaled the scores of each question item to determine the value of satisfaction. The survey questions are listed in Table 1, and in this experiment, participants answered the five questions in Table 1 on a 7-point scale (ranging from 1 to 7), and the sum of the scores for the pair of participants indicated satisfaction. Due to the influence of students’ characteristics such as effort and motivation on...
course satisfaction, Uchida used satisfaction as an evaluation metric for cooperative learning, as it is believed to be useful in demonstrating the relationship between the pair effect and personality traits. The satisfaction for participants \(a\) and \(b\) was denoted by \(Z_{a,b}\). When the pairs were matched as (participant 0 and participant 1), (participant 2 and participant 3), (participant 4 and participant 5), (participant 6 and participant 7), (participant 8 and participant 9), and (participant 10 and participant 11), the resulting data set of satisfaction was \(\{Z_{0,1}, Z_{2,3}, Z_{4,5}, Z_{6,7}, Z_{8,9}, Z_{10,11}\}\).

3.4 An Analysis of the Relationship Between Emotion Recognition Results and Evaluation of Pair Work

Let \(X_p\) be the array of emotional levels of participant \(p\) as represented in equation (2). When matching participants 0 through 11 as pairs (participant 0 and participant 1), (participant 2 and participant 3), (participant 4 and participant 5), (participant 6 and participant 7), (participant 8 and participant 9), and (participant 10 and participant 11), let \(\{|A_0 - A_1|, |A_2 - A_3|, |A_4 - A_5|, |A_6 - A_7|, |A_8 - A_9|, |A_{10} - A_{11}|\}\) be the dataset of differences in neutral emotional levels between pairs, and let \(\{A_0 + A_1, A_2 + A_3, A_4 + A_5, A_6 + A_7, A_8 + A_9, A_{10} + A_{11}\}\) be the dataset of the sums of neutral emotional levels between pairs. Similarly, define datasets for the differences and sums of seven other emotional states (calm, happy, sad, angry, fearful, disgusted, surprised), as well as for the differences and sums of the eight emotion recognition results, and for the datasets of activation and satisfaction (giving 32 combinations in total).

![Graphs showing relationships between activations and differences](image)

**Figure 4.** Relationship between Activations and Differences in Angry Emotional Levels (upper left), Disgusted Emotional Levels (upper right), and Surprised Emotional Levels (lower left). Relationship between Satisfaction and Sum of Fearful Emotional Levels (lower right).

In the analysis, we conducted a test of non-correlation, with the null hypothesis that “there is no correlation between the two variables” and the alternative hypothesis that “there is a difference between the population mean and the sample mean.” We set the significance level to 0.05 and compared test statistic \(t\), calculated from the correlation coefficient and a \(t\)-distribution value of \(k = 2.7\). For all 32 combinations, we calculated test statistic \(t\) and found that the null hypothesis was rejected in four cases: the relationship between the dataset of differences in angry emotional levels and the dataset of activation \((t = 3.3 > k)\), the relationship between the dataset of differences in disgusted emotional levels and the dataset
of activation ($t = 4.7 > k$), the relationship between the dataset of differences in surprised emotional levels and the dataset of activation ($t = 3.5 > k$), and the relationship between the dataset of sums of fearful emotional levels and the dataset of satisfaction ($t = 3.4 > k$). Therefore, we can say that the correlation coefficients between the dataset of differences in angry emotional levels and the dataset of activation ($-0.86$), between the dataset of differences in disgusted emotional levels and the dataset of activation (0.92), between the dataset of differences in surprised emotional levels and the dataset of activation (0.87), and between the dataset of sums of fearful emotional levels and the dataset of satisfaction, ($-0.86$) are significant at the 0.05 level. Graphs showing each relationship and the regression line created by the least squares method are shown in Figure 4. Assuming that the datasets for the differences in angry, disgusted, surprised, and the sum of fearful emotional levels are independent, we can consider the following four points in our discussion.

- The smaller the difference in the angry emotional level, the greater the activation.
- The larger the difference in the disgusted emotional level, the greater the activation.
- The larger the difference in the surprised emotional level, the greater the activation.
- The smaller the sum of the fearful emotional level, the greater the satisfaction.

### 3.5 The Development of a Matching Algorithm

After considering four aspects related to emotional levels, we developed a matching algorithm. The procedure for developing the algorithm is explained as follows:

First, we determined pairs of two individuals ($P_1$, $P_2$, $P_3$, $P_4$) from all the learners who wanted to be matched. $P_1$ indicates the point that improves as the difference in the angry emotional level becomes smaller, $P_2$ indicates the point that improves as the difference in the disgusted emotional level becomes larger, $P_3$ indicates the point that improves as the difference in the surprised emotional level becomes larger, and $P_4$ indicates the point that improves as the sum of the fearful emotional level becomes smaller. If we assume that the array of emotional levels for participant $p$ is expressed as $X_p$, then the combinations of $P_1$, $P_2$, $P_3$, and $P_4$ for learners $a$ and $b$ are defined as expressions (3), (4), (5), and (6), respectively.

Next, assuming that the data groups for the difference in the angry emotional level, the difference in the disgusted emotional level, the difference in the surprised emotional level, and the sum of the fearful emotional level were all independent, we normalized these four data groups to have a minimum of 0 and a maximum of 1 to treat these four points equally. When $P_{1,a,b}$, $P_{2,a,b}$, $P_{3,a,b}$, and $P_{4,a,b}$ were denoted as the normalized data, $P_{1,a,b}'$, $P_{2,a,b}'$, $P_{3,a,b}'$, and $P_{4,a,b}'$, respectively, the overall point $P$ for learners $a$ and $b$ was defined as expression (7).

Finally, as we contend that higher $P$ values correspond to increased pair work activity and satisfaction, we recommend learners as a match in the order of their $P$ values. For example, if we calculate the points such that $P_{0.1} < P_{0.2} < P_{0.3}$ among learners 0, 1, 2, and 3, we recommend learners 3, 2, and 1 in that order to learner 0.

$$P_{1,a,b} = \frac{1}{|E_a - E_b|}$$  \hspace{1cm} (3)

$$P_{2,a,b} = |G_a - G_b|$$ \hspace{1cm} (4)

$$P_{3,a,b} = |H_a - H_b|$$ \hspace{1cm} (5)

$$P_{4,a,b} = \frac{1}{F_a + F_b}$$ \hspace{1cm} (6)

$$P_{a,b} = P_{1,a,b}' + P_{2,a,b}' + P_{3,a,b}' + P_{4,a,b}'$$ \hspace{1cm} (7)
4. Evaluation of a Matching System

An evaluation experiment was conducted on the proposed matching system. The participants consisted of a total of 10 university and graduate students (5 male and 5 female). The proposed method was evaluated by comparing it with a pairing method that used sex and basic academic skill, which has been confirmed to be effective by Uchida (Uchida, 2016). Uchida explains the pairing method using sex and basic academic skill as follows:

(1) Sort the students in descending order based on their basic academic skill scores.
(2) Pair them according to the sorting results.
(3) If the pairing results are in same-sex pairs in (2), replace them with opposite-sex pairs with a small difference in basic academic skill scores. However, if the difference in basic academic skill scores exceeds 2 out of 20 points, they will be kept as same-sex pairs.

An experiment was conducted to compare the existing method of matching using characteristics obtained from surveys or tests as indicators with the proposed method of matching using the heterogeneity related to learners’ characteristics as indicators.

4.1 Experimental Methods

First, the experimental participants learned with partners matched by using the existing method. In Uchida’s experiment, a 20-point test was conducted to measure basic academic skills. In this experiment, a programming test with a maximum score of 20 points was conducted to measure basic programming skills in robot programming. Then, based on Uchida’s algorithm, 10 participants were matched and engaged in pair learning, after which pair work was evaluated using Uchida’s activation and satisfaction.

Next, the participants learned with partners matched using the proposed method. All participants learned with virtual partners, and the speech data obtained were analyzed to match participants based on the matching algorithm developed. Ten participants were matched and engaged in pair learning, after which pair work was evaluated in the same manner. Finally, the pair work evaluation of the existing method and the proposed method were compared.

4.2 Experimental Results and Discussion

4.2.1 Comparison of the mean values of pair work evaluation

Figure 5. Comparison of Existing and Proposed Methods (MT) for Activation (left) and Satisfaction (right).

Figure 5 compares the activation and satisfaction of the existing method and the proposed method. The mean value of activation for the existing method was 0.21 and for the proposed method it was 0.18. Assuming that the pairs in the existing and proposed methods were independent since the participants were the same but the combinations were different, we
performed an analysis to determine whether there was a difference in the mean values of pair work evaluation for both methods. A two-sided t-test without correspondence was performed for the analysis. The null hypothesis was that “there is no difference in the mean values of pair work evaluation between the existing and proposed methods,” and the alternative hypothesis was that “there is a difference in the mean values of pair work evaluation between the existing and proposed methods.” We compared the t-value with the boundary value of \( k = 2.2 \) for a two-sided test with a significance level of 0.05. As a result, the null hypothesis was not rejected, and it was found that there was no significant difference in the mean values of pair work evaluation between the existing and proposed methods, both for activation \((t = 0.83, p = 0.42)\) and satisfaction \((t = -0.54, p = 0.60)\).

4.2.2 Comparison of the mean values of pair work evaluation

To investigate the effectiveness of the proposed method over existing approaches for specific types of participants, we analyzed individual satisfaction (on a scale of 35 points). First, we examined the participants’ basic academic skills (based on pre-test results), their satisfaction when matched and learning with partners using the existing method, and their satisfaction when matched and learning with partners using the proposed method. The differences in satisfaction (calculated by subtracting the satisfaction obtained with the existing method from that of the proposed method) are presented in Table 2, where participants with lower basic academic skills showed higher satisfaction with the proposed method compared to the existing approach. Consequently, we proceeded with an analysis to explore the relationship between basic academic skill scores and differences in satisfaction.

In the analysis, we conducted a test of non-correlation, with the null hypothesis that “there is no correlation between the two variables” and the alternative hypothesis that “there is a difference between the population mean and the sample mean.” We set the significance level to 0.05 and compared the test statistic \( t \), calculated from the correlation coefficient and a \( t \)-distribution value of \( k = 2.7 \). As a result, since the null hypothesis was rejected \((t = 3.3 > k)\), we can assert that the correlation coefficient \((-0.84)\) between academic skill scores and differences in satisfaction is significant in terms of \( t \). The relationship and the regression line created by the least squares method are shown in Figure 6. From there, it can be contended that the proposed matching system may have a greater potential to yield positive effects for learners with lower basic academic skills.

Table 2. Personal Comparison of Satisfaction with the Existing and Proposed Methods (MT).

<table>
<thead>
<tr>
<th>Participant</th>
<th>Academic Skill Score</th>
<th>Satisfaction (Existing MT)</th>
<th>Satisfaction (Proposed MT)</th>
<th>Difference in Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>29</td>
<td>27</td>
<td>-2</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>29</td>
<td>26</td>
<td>-3</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>34</td>
<td>35</td>
<td>+1</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>30</td>
<td>33</td>
<td>+3</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>28</td>
<td>34</td>
<td>+6</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>20</td>
<td>29</td>
<td>+9</td>
</tr>
</tbody>
</table>

4.2.3 Direct evaluation through questionnaires allocated to participants

In addition to the satisfaction survey regarding pair work, we conducted interviews after all the experiments to ask which learning partner the participants found easier to study with and the reasons for their preferences. During the interviews, we deliberately concealed whether the partners were matched using the existing method or the proposed method. The results of the survey are presented in Figure 7. Among the participants, two individuals were matched
with the same partners using both the existing and proposed methods, so we did not obtain survey responses from them. However, 70% of the participants stated that they found it easier to learn with partners matched through the existing method.

Furthermore, in the interviews with participants who rated the proposed method more favorably, the majority reported that their learning partners in the proposed method were easier to talk to, making it simpler to ask questions and seek advice. However, some participants mentioned that during the proposed method period, as it was their second class together, they felt less resistance in interacting with their partners compared to the first class with the existing method, possibly because they were getting accustomed to the experimental setup.

These results suggest that the order of participation and the nature of the tasks may have introduced bias among the participants. To support the statement made in section 4.2.2, that “the proposed matching system may have a greater potential to produce positive effects for learners with lower basic academic skills,” further experiments are needed. These experiments should include increasing the number of participants and ensure no bias by assigning different groups of participants to learn with either the existing or the proposed method.

![Graph showing the relationship between Basic Academic Skill Score and Differences in Satisfaction.](image)

*Figure 6. Relationship between Basic Academic Skill Scores and Differences in Satisfaction.*

![Pie chart showing the percentage of participants who found learning partners easy to learn with.](image)

*Figure 7. Survey results on “Learning partner that participants in the experiment found most easy to learn with.”*

### 5. Conclusion

In this study, we proposed a matching system for automating group formation by the heterogeneity of emotion recognition results in online collaborative learning during paired robot programming. We conducted an evaluation experiment comparing the existing and proposed methods of the matching system to clarify whether group formation based on the heterogeneity of emotion recognition results is more effective than using conventional methods. While no significant difference was observed in the average pairwise evaluation ratings, we found that the proposed method resulted in higher satisfaction ratings for those
participants with lower basic skills. However, the experiment had limitations such as a small number of participants and thus insufficient data, and the possibility of bias due to the comparison of the existing and proposed methods with the same participants and different task content.

Therefore, in future studies, we plan to increase the number of participants and conduct sufficient testing to refine the matching algorithm. In addition, we plan to conduct between-participant comparisons with standardized task content, rather than within-participant comparisons. In this way, we will carefully account for biases due to differences in order of participation and task content, and continue with the evaluation experiments.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H01112.

References


Assessing college students’ sense of community for advancing community knowledge

Chih-Hui SEETa*, Yi-Ning TSAIb*, Guo-Tsai HUNGc & Huang-Yao HONGd

a,b,dDepartment of Education, National Chengchi University, Taiwan
bDepartment of Japanese Studies National Taichung University of Science and Technology, Taichung
*111152518@g.nccu.edu.tw, *108152513@nccu.edu.tw, barryhuck@gmail.com, hyhong@nccu.edu.tw

Abstract: This case study aims to examine students’ sense of community (SOC) during online learning. Participants were college students data came from the “Classroom Community Scale” and students’ online discussion in a platform designed based on knowledge building theory and pedagogy. The results showed that students were able to develop a strong sense of community and perceived learning as a knowledge creating endeavor as guided by knowledge building pedagogy.

Keywords: sense of community, knowledge building, online discussion

1. Introduction
Knowledge economy has accelerated the speed of educational innovation and reform that values collaborative knowledge creation more than individual knowledge acquisition. Knowledge creation or knowledge building (KB) is defined as a social process focusing on sustained idea improvement and on knowledge advancement as a collective endeavor. In a KB environment, students continually contribute, refine, and build on their ideas for advancing collective knowledge as a knowledge community (Bereiter & Scardamalia, 2014). Knowledge building is thus idea-centered (Scardamalia & Bereiter, 2006) and its community particularly highlights the importance of working creatively and interactively with ideas (Hong & Scardamalia, 2014), and members of the community are deemed as legitimate knowledge contributors for the overall achievement of group knowledge goals. In a knowledge building community, how to develop a stronger sense of community (SOC), is important and remains a challenge, which has also received increasing attention from many educators. With stronger SOC, people can feel that they belong to a group or the greater community and those who are in the group or in the community can also better relate to each other with similar ideas or learning commitment. Stronger SOC also implies that community members can be more easily motivated to interact with other community members for sustained knowledge work that leads to the achievement of the community goals (Cacciamani & Perrucci, 2020). With increasing SOC, community members may also feel more psychological safe and thus are more likely to participate more in a more proactive manner. For students in a knowledge building community, SOC is therefore essential as it can foster a better class climate for collective knowledge creation. Accordingly, in this study, we engage students in KB activities, hoping that through various guided KB activities, students could enhance their SOC as a knowledge-based community.

2. Method
This case study adopted a mixed-methods research design to explore students’ sense of community in a knowledge building environment. Participants were 21 college and postgraduate students taking a class titled “Study on Instructional Design of Educational Technology” in a national university in Taiwan. The course encouraged students to interact with one another and to contribute ideas to improve their assigned design projects. Two sets of surveys were administrated. One is a pre-established “Classroom Community Scale” (Rovia, A. P., 2020) which evaluates students’ sense of community (in two dimensions
including awareness of connectedness and awareness of learning); and the other is “Types of Knowledge Community,” which contains self-develop, open-ended questions to assess students’ sense about three types of knowledge community (i.e., knowledge telling, sharing, and creating communities). Sample questions are such as: What is a knowledge community? What constitutes a good knowledge community? Data were mainly collected from online activities/discussion, and the above mentioned two surveys. A paired-sample t-test analysis was conducted to examine students’ perceived sense of community (survey 1) and sense of knowledge community (survey 2). Z-test was executed to classify students’ perception of knowledge community as knowledge-telling, knowledge-sharing, or knowledge-creating community.

3. Findings
As Figure 1 shows, about students’ sense of community, first it was found the mean number of times from the pre-test to the post-test in the awareness of connectedness were from 2.88 to 3.46 (t=3.58***, p=.001). It shows that there were significant differences after one-semester’s KB activity and students enhanced their sense of community in a knowledge building community. On the other hand, Figure 2, about students’ sense of 3 types of knowledge community, it was found the mean number of times from the pre-test to the post-test in knowledge telling were from 1.89 to 0.58 (t=-3.842***). It shows that while there was no statistically significant difference in students’ perception of knowledge sharing community and knowledge creating community, there was significant differences in the pre and post-test of students’ perception of knowledge telling community. Figures 3 further shows social network analysis. It was found that interactions regarding students’ ideas submitting, exchanging, and suggestions and elaborations for their design projects are very frequent and consistent. Finally, regarding types of knowledge community, it also indicates that students were more aware of the importance of knowledge community for knowledge creation in a KB community. Figure 4 also shows students weekly online activity performance.

![Figure 1. Students’ sense of community in KB community](image1)

![Figure 2. Types of knowledge community](image2)

4. Discussion
This study investigated students’ SOC after engaging students in an online knowledge building environment. It was found that first, the participants enhanced their SOC, both on the dimensions of “connectedness” and “learning”; second, regarding students’ sense of 3 types of knowledge community, students tended to consider it as knowledge telling before the course but they changed to that of knowledge creating after the whole semester. To conclude, students enhance their SOC via KB and students were able to become more connected and interactive community to achieve community's learning goals. In sum, our study shows that working in a KB community makes it possible for student’s developing a stronger sense of knowledge community (Bereiter & Scardamalia, 2014). While knowledge creation is meaningful for the future progress of human beings, SOC is useful for enhancing the goodwill to work collaboratively among members in a society that values group work more than...
individual work. Although this study shows that SOC can be fostered through a KB learning environment, future discussion from different educational context and researches are needed to triangulate the findings in this study.

Figure 3. The author networks

Figure 4. Online weekly performance in KF

Each student’s weekly-based action on KF

References


Rovia, A. P. (2020). Development of an instrument to measure classroom community. The Internet and Higher Education. 5, 197-211.

Teachers developing more creative learning views via online knowledge building activities
Mei-Ju Chen*, Chao-Yu Guo**, Huang-Yao Hong**
meiju@mx.nthu.edu.tw, chaoyu@nccu.edu.tw, hyhong@nccu.edu.tw
National Tsing Hua University*, National Chengchi University**

Abstract: This study investigates how teachers who were engaged in online knowledge building via guided perspective-taking (PT) activities to discuss their critical teaching incidents (CTIs) helped them to develop creative learning views that see learning not merely as knowledge acquisition and exchange, but also as knowledge creation. Data were mainly from online discussion. The main findings indicate that PT-based KB activities in a technological platform can guide teachers to reflect deeply on the CTIs they experienced before and on how they might be able to help students learn better in a more constructivist and creative manner.

Keywords: perspective taking, knowledge building, online learning environment, discussion

1. Introduction

The importance of helping teachers develop more informed, creative learning views, so that they may be more likely to succeed in engaging their students’ in more constructive and innovative learning has been well recognized in literatures (e.g., Antink-Meyer & Arias, 2022; Park et al., 2022). How to engage teachers in in-depth reflection on their views of learning, however, remains a pedagogical challenge for many educators. In the present study, we introduce perspective taking (PT) as a means to engage teachers in such deep reflection. PT is a way of understanding one’s own mind or other’s minds from inside (Kloo et al., 2021), and it can help teachers to develop self-awareness and emotional clarity (Campbell & Pennebaker, 2003). In education, most PT-related research is concerned with students (Kazuo et al., 2020), focusing on how teachers use PT to help students to develop multiple views or cultural understanding (Abacioglu et al., 2020; Athanases & Sanchez, 2020). Yet, PT is rarely used as a tool to improve teacher professional development (TPD). As a reflective means, PT can provide three different introspective levels to help teachers conduct progressively deeper reflection, from self-reflection (I–PT); to self-distancing reflection (meaning to reflect upon personal teaching experiences through the perspective of an outside observer; Other–PT); to more refined self-reflection (Back to I–PT) via interpreting and synthesizing insights from previous I-PT and Other-PT reflection. In particular, when the object of reflection is on critical teaching incidents (CTIs), PT may reveal many unknown aspects relating to teachers’ self-understanding of own teaching practice (Mor-Hagani & Barzilai, 2022) and to their teaching emotions (Rademacher et al., 2010). Therefore, CTIs as recurring teaching problems or challenges (Joshi, 2018) are appropriate topics for teachers’ reflection (Badia et al., 2021) to help teachers develop robust reflection-in-action capacity (Schon, 1983). To this end, we employed PT in an online KB community (Hong & Scardamalia, 2014) to foster both individual and group reflection among teachers to help them build more in-depth self-knowledge of teaching (as contrasted with the typical core teaching knowledge or skills specified in teacher education curriculum).

The purpose of the study was to explore how in-service teachers can collaboratively help one another to reflect on their CTIs via PT in an online platform for developing more informed and productive learning views that can help them become more mentally adaptive teachers. Building on the above discussion, the main research question is: Does engaging participants in PT while discussing critical teaching incidents (CTIs) in an online knowledge building environment help them develop more informed learning views to help their students learn more constructively?

2. Method

We recruited fifty-one incumbent teachers as participants who were taking a graduate course in a university from Taiwan. We engaged these participants in online learning in a knowledge building
environment called Knowledge Forum (KF). We then introduced PT as a reflective measure to deepen participants’ discussion about some critical teaching incidents (CTIs) occurred during their recent teaching. The course duration was 18 weeks which were intentionally segmented into three phases/stages for PT activities—starting from I-PT, to Other-PT, and to Back-to-I-PT (see Figure 1 for our pedagogical design). Data sources and analyses focus on (1) the number of notes contributed, read, and modified in KF, and words written per note; (2) discussion that reveals participants’ learning views and emotions (in this paper we did not report the emotional aspect). We adopt Hong’s (2009) conceptualization of learning as acquisition, or as participation, or as creation into our coding scheme and Kappa coefficient of inter-coder reliability was calculated to be 0.74. Wilcoxon signed-rank test was employed for analysis.

Figure 1. Students working in Knowledge Forum to perform three different perspective taking activities

3. Findings

Table 1 summarizes the key findings from the analysis throughout the whole semester. First, regarding the overall online reflection trend, our analysis indicates that there was a decreasing trend of note contributing, note modification, and note reading from Stage 1 to 3. However, when looked into the number of words (i.e., Chinese characters) written per note, there was an increasing trend from Stage 1 to Stage 3 (Figure 2). The fact that participants gradually posted fewer notes but wrote more words in each note implies that participants progressively became more reflective on the CTIs discussed online in KF from Stage 1 to Stage 3.

Second, regarding more specific reflection trend on learning views, Z tests to compare Stages 1 and 3 indicated that there was a significant decrease in participants’ reflection on acquisition-oriented views ($z = 5.064^{**}$, $p < 0.01$) and a significant increase in participants’ reflection on creation-oriented views ($z = -2.428^*$, $p < 0.05$; see Figure 3), indicating that participants were able to develop more informed learning views after taking this course.

Table 1. Changes in Participants’ Overall Online Reflective Activities and Specific Reflective trend on Views of Learning in 3 PT stages

<table>
<thead>
<tr>
<th>Change in Reflection</th>
<th>Knowledge building via Guided Perspective-Taking (PT)</th>
<th>Z Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stage 1</td>
<td>Stage 2</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>$M$</td>
</tr>
<tr>
<td><strong>Online activities</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contribution</td>
<td>179</td>
<td>3.5</td>
</tr>
<tr>
<td>Modification</td>
<td>351</td>
<td>6.9</td>
</tr>
<tr>
<td>Reading</td>
<td>1,734</td>
<td>34</td>
</tr>
<tr>
<td>Words per note</td>
<td>10,311</td>
<td>202</td>
</tr>
<tr>
<td><strong>Views of learning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisition</td>
<td>23 (21.70%)</td>
<td>0.45</td>
</tr>
<tr>
<td>Participation</td>
<td>39 (36.79%)</td>
<td>0.76</td>
</tr>
<tr>
<td>Creation</td>
<td>44 (41.51%)</td>
<td>0.86</td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Note: contribution excluding feedback notes.
Discussion

This study engaged the participating in-service teaching in three different perspective-taking (PT) activities (including I-PT, Other-PT, and Back-to-I-PT) while discussing some critical teaching incidents (CTIs) in an online knowledge building environment (i.e. Knowledge Forum, KF). Conventional teachers’ professional development tends to focus on helping teachers develop core teaching knowledge and skills. In contrast, our study is trying to guide teachers to perform self and social reflection on their learning views. The main finding indicates that discussing critical teaching incidents via taking different perspectives indeed can effectively help teachers become more reflective to come up with more ideas for how to teach better, or to help students learn better. Accordingly, they were more willing to embrace a more informed and creative learning view. Thus, we conclude that PT as a self-distancing, reflective means, indeed, can transform teachers’ past teaching experience into more practical self-knowledge that is essential for more effective teacher professional development. At the same time, our finding also suggests that KF was able to provide the participants an online knowledge space for perform the required self and social (collective) reflection (Kim, et al., 2018). It also helps the participating teachers to form some community-based, in addition to merely individual-based, professional development experience. Future analysis will focus on analyzing the emotional impact from PT activities and on how emotional transformation may be related to the change in participants’ learning views.

References


Unveiling University Students' Data Literacy: A Case Study on Modeling Reasoning in Data Mining Projects

Tianqi ZHANG
East China Normal University, China
51214108044@stu.ecnu.edu.cn

Abstract: In the era of big data, cultivating students' data literacy is of paramount importance. Data literacy encompasses the abilities to collect, manage, analyze, and apply data. As data science is a superset composed of mathematics and statistics, computer science, and specific application fields, it is crucial to investigate data literacy development through the lens of statistical education. An essential component of statistical reasoning is modeling reasoning, which is also fundamental to data literacy. However, limited research exists on the manifestation of data literacy among university students in real-world tasks. Therefore, this study explores the forms of modeling reasoning exhibited by university students in a data mining project. The findings reveal that, in a real teaching setting, university students' modeling reasoning ability follows a spiral progression. Considering these insights, modeling reasoning, as the core of data mining activities, plays a pivotal role in fostering students' data literacy. To cultivate data literacy among university students in the age of big data, we recommend implementing project-based learning, incorporating ill-structured problems and real, complex, massive datasets as project backgrounds.

Keywords: Data literacy, modeling reasoning, statistical reasoning, project-based learning

1. Introduction

In the era of big data, the rapid advancement of intelligent technology and its increasing integration into everyday life have transformed data into a valuable resource in the 21st century. Both society and the labor market now demand graduates who possess the knowledge and skills to effectively utilize statistics, data management, and computer science to make informed decisions. Consequently, the cultivation of data literacy, which encompasses the abilities to collect, manage, analyze, and apply data, has become increasingly crucial. In fact, data literacy has emerged as an upgraded version of information literacy within the realm of 21st-century skills. To address the growing demand for data experts in the age of big data, universities have established data science majors. Researchers argue that data science should encompass mathematics and statistics, computer science and programming, as well as specific application fields (Rosenberg, Lawson, Anderson, Jones & Rutherford, 2019). As data availability improves and technological advancements introduce user-friendly data processing software, researchers in statistical education are actively exploring methods to foster students' data literacy in the era of big data (Ben-Zvi, Makar & Garfield, 2018).

Within statistical education, modeling reasoning is a fundamental underlying ability in both statistical reasoning and data literacy (Bakker & Hoffmann, 2005; Wild & Pfannkuch, 1999). Rubin (2019) synthesized key aspects derived from a series of works related to data processing in the era of big data, including context, variation, aggregation, visualization, and inference. In particular, inference, specifically Informal Statistical Inference (ISI), is closely associated with modeling inference. In the realm of statistical education, modeling reasoning serves as the foundation for data analysis and result interpretation. It involves simplifying
phenomena based on data and theories, generating models that can explain, predict, or define these phenomena (Aridor & Ben-Zvi, 2017).

Modeling reasoning not only simplifies the process of defining, explaining, and predicting phenomena based on data and theories (Aridor & Ben-Zvi, 2017), but it also constitutes a key element of data literacy and statistical reasoning (Wild & Pfannkuch, 1999). Generally, models are regarded as forms of interpretation, while modeling is a process that simplifies, evaluates, and enhances our understanding of a phenomenon using key theories and data from a specific discipline, either by incorporating it into existing theoretical frameworks or by generating new discoveries (Lehrer & Schauoble, 2010). Models can be classified as abstract models (conceptual models) or concrete models (e.g., figures and tables). Abstract models represent real-world systems and conjecture about their behavior to describe, explain, predict, and elaborate on their behavior (Wild & Pfannkuch, 1999).

Concrete models, on the other hand, serve as tools to represent processes such as identifying key components or properties of a population, making predictions or samples, and drawing inferences about the representativeness of a random sample (Aridor & Ben-Zvi, 2017). Modeling is a multifaceted process in which the role of the model evolves alongside changes in thinking (Gravemeijer, 1999). Initially, a model takes the form of an "informal reasoning model." As a new concept emerges, the model's role changes, and it transforms into a "formal reasoning model." These two processes are accompanied by a third process that shapes the model into a set of symbols guiding previous reasoning processes (Gravemeijer, 1999).

Practical activities in statistical education are often seen as a form of modeling as they involve understanding variation and uncertainty, processing data, and constructing models (Lehrer & English, 2018). The modeling process entails evaluating and optimizing models, including generating new theoretical ideas or discoveries based on data (Dvir & Ben-Zvi, 2018). Modeling reasoning can be viewed as an analogical process that simplifies real phenomena, elucidates connections and relationships among their components, and deals with inherent uncertainties (Wild & Pfannkuch, 1999). Familiarity with the modeling reasoning process greatly aids the cultivation of data literacy. In an effort to enhance the quality of statistical education in primary schools, Biehler et al. interviewed four pre-service primary school teachers who participated in a statistical modeling reasoning training course. They evaluated and analyzed the teachers' reasoning processes, scrutinized their utilization of Tinkerplots to model statistical situations, and assessed their ability to evaluate models in accordance with the given statistical situations, providing recommendations accordingly (Biehler, Frischemeier & Podworny, 2017). The learning environment associated with modeling reasoning can also facilitate the development of students' data literacy. Conway et al. conducted quasi-experimental research on primary school students and found that adhering to the principles of a statistical reasoning learning environment in both beliefs and practices had a more positive impact on students' statistical reasoning ability compared to traditional classroom teaching (Conway, Martin, Strutchens, Kraska & Huang, 2019). This indicates that training in modeling reasoning is pivotal in statistical and data science teaching activities, and that modeling reasoning ability plays a vital role in students' comprehension, analysis, and application of data.

Despite the considerable research on modeling reasoning in primary and middle school students, few studies have focused on the data literacy of university students (Setiawan & Sukoco, 2021). Currently, there is a dearth of empirical research examining the modeling reasoning ability of university students. Thus, a better understanding of how university students manifest modeling reasoning ability is crucial for the development of data science programs. Additionally, as there is limited guidance on the theoretical exploration and teaching practices for specific data literacy training, further theoretical and empirical research is necessary to guide the development and implementation of data science education.

Therefore, this study aims to explore the manifestation of data literacy among university students during real-world tasks, specifically by examining the forms of modeling reasoning employed by university students in a data mining project. Due to the limited knowledge regarding modeling reasoning among university students, a qualitative research
approach, specifically a case study, was deemed appropriate for exploring this phenomenon. The study utilized a bottom-up grounded theory method to provide a detailed analysis of the manifestation of modeling reasoning among university students.

2. Method

2.1 Participants

The participants in this study were three university students who enrolled in an educational data mining course at a university in Shanghai, China. These students worked collaboratively on a data mining project as part of the course. The project involved analyzing real and complex education data to predict students' answering efficiency in a specific scenario. The scenario was as follows: Some students took an online math test consisting of two sets of exams (exam A and exam B) with the same type and number of questions. Each test lasted for 30 minutes, and once the time expired, students were automatically terminated and unable to continue. The project provided students with four data documents: (1) a sequence of students’ 30-minute behavior on exam A for Group 1 students, (2) the efficiency of Group 1 students on exam B, (3) the performance of Group 2 students on exam A, and (4) the prediction results of the answering efficiency of Group 2 students on exam B. The objective of the project was for students to build a model based on the performance of Group 1 students on exam A and their efficiency in completing exam B, and then use this model to predict the efficiency of Group 2 students in completing exam B based on their performance on exam A.

The dependent variable in this project was the answering efficiency of exam B, which was defined as a binary variable: “Yes” for efficient and “No” for inefficient. Efficiency was determined based on two criteria: (1) whether all questions on exam B were answered and (2) whether the time allocation for each question was reasonable. Reasonableness was determined by comparing the time distribution of all students for each question and setting the threshold as the shortest possible time for each question (specifically, the first 5% of the time distribution).

The independent variables included student ID, test type, a unique question number for each question, question type, specific behaviors of students during the answering process, additional information about their behaviors, and the time stamp of each answering behavior.

2.2 Procedure

Throughout the 9-week course, the students utilized the Zoom meeting tool to engage in discussions twice a week, with each session lasting 50-60 minutes. The entire discussion process was recorded using the Zoom meeting feature. These recorded videos were manually transcribed to obtain a written record of each discussion.

2.3 Data Analysis

The transcribed discussions resulted in a conversation document comprising a total of 82,003 words and 3,209 statements. A statement was defined as a complete conversation in which a student participated and contained at least one sentence.

The data analysis employed a bottom-up grounded theory method to derive a comprehensive understanding of the data. Initially, the conversation document was analyzed sentence by sentence using open coding to identify and extract several categories. Subsequently, specific categories were grouped together through axial coding to generate overarching themes. Finally, each category was further refined using selective coding to extract core concepts and capture the key content of the conversations.

To ensure the reliability of the coding and analysis, two researchers coded respectively under the guidance of professionals and got good consistency (Kappa=0.81). Furthermore,
data from different sources, such as the discussion videos and the team’s final project report, were also integrated into the category analysis. This additional data served as a cross-reference and validation for the coding process.

3. Findings

3.1 Overall Status

To depict the manifestation of university students, the coding results are shown in Table 1. By coding the conversation document sentence by sentence, and with repeated reading, comparison and modification, the whole conversation document was finally divided into 41 activity segments, each representing a complete activity discussion among the students in the group, such as understanding the task requirements and dealing with missing values. After removing repeated activities, 19 first-level codes were identified through open coding (see Table 1). Axial coding was then conducted to analyze the categories based on the 19 first-level codes and establish relationships between them. The first-level codes were organized chronologically, and their relationships in terms of causality, situation, similarity, difference, function, and process were considered, resulting in the formation of 8 second-level codes (see Table 1). Finally, selective coding was performed, and all categories were classified into corresponding core categories that encompassed all the activities in the session document. Five core categories, referred to as third-level codes, were extracted: understanding tasks, organizing data, selecting variables, selecting sampling methods, and selecting modeling methods (see Table 1).

Table 1. Coding Results of Bottom-Up Grounded Theory

<table>
<thead>
<tr>
<th>First-level Codes</th>
<th>Second-level Codes</th>
<th>Third-level Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding the task</td>
<td>Understanding the task</td>
<td>Understanding the task</td>
</tr>
<tr>
<td>Assigning phased task</td>
<td>Preparing software environment</td>
<td>Preparing software environment</td>
</tr>
<tr>
<td>Dealing with missing values</td>
<td>Preprocessing data</td>
<td>Organizing data</td>
</tr>
<tr>
<td>Exploring preprocessing data</td>
<td>Reflecting on preprocessing data</td>
<td>Processing data</td>
</tr>
<tr>
<td>Preprocessing data</td>
<td>Exploring understanding data</td>
<td>Preprocessing data</td>
</tr>
<tr>
<td>Reflecting on understanding data</td>
<td>Exploring independent variables</td>
<td>Selecting variables</td>
</tr>
<tr>
<td>Selecting independent variables</td>
<td>Reflecting on independent variables selection</td>
<td>Selecting independent variables</td>
</tr>
<tr>
<td>Exploring sampling methods</td>
<td>Understanding sampling method</td>
<td>Selecting sampling method</td>
</tr>
<tr>
<td>Understanding sampling</td>
<td>Reflecting on sampling methods</td>
<td>Sampling</td>
</tr>
<tr>
<td>Exploring modeling process</td>
<td>Exploring modeling process</td>
<td>Modeling</td>
</tr>
<tr>
<td>Reflecting on modeling process</td>
<td>Understanding modeling process</td>
<td>Modeling</td>
</tr>
<tr>
<td>Selecting model</td>
<td>Evaluating model</td>
<td>Selecting model</td>
</tr>
<tr>
<td>Optimizing model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

By summarizing the second-level and third-level codes, it was observed that the modeling reasoning process of university students in the educational data mining project followed a spiral pattern. This process involved the steps of defining the problem, generating
and selecting variables and attributes, selecting modeling and sampling methods, organizing and structuring data, exploring and analyzing data, interpreting data, and presenting the results (see Figure 1). The process went through several iterations, with the students continuously optimizing and improving their understanding of specific concepts, selection of independent variables, data processing methods, sampling techniques, and modeling methods over time, ultimately leading to a progressive spiral form.

3.2 Detailed Manifestation

Since the goal of the data mining project was to build a prediction model for students’ answer efficiency, the entire process of completing the project represented a comprehensive modeling reasoning process. Additionally, since the students engaged in discussions twice a week, it was evident from the analysis that they would review the task flow at the beginning of almost every weekly discussion and allocate tasks accordingly. This iterative nature of their discussions contributed to the spiral development of their modeling reasoning process, wherein each week built upon the progress made in the previous week. Therefore, to demonstrate the spiraling process of modeling reasoning, relevant conversation fragments that were important and non-repetitive were selected based on the chronological order of task completion.

3.2.1 Understanding the meaning of the task

After reading the task description document, the students (referred to as student X, student Y, and student Z) engaged in a discussion to accurately comprehend the task's purpose, define its objectives, and determine the analytical process and steps [2-104] (Numbers in square brackets denote the statement numbers of the conversation document). Since the process of understanding the meaning of the task was relatively fragmented, and their understanding of the task's purpose did not undergo significant changes thereafter, the project report is used as a demonstration. The students clarified the task’s objectives as follows: “How to predict the answering efficiency of (another) group of students in exam B by analyzing the data of the existing group of students’ answering behavior in exam A and the answering efficiency in exam B, with the ultimate goal of maximizing the accuracy of the prediction” [Project report, p1] (means the source of information is from page 1 of the team project report). They also depicted the analysis process in Figure 2 [Project report, p2].

![Figure 1. The spiral progress of modeling reasoning](image)

Although the students had a clear understanding of the general process, they encountered some difficulties and uncertainties in the execution and comprehension of specific steps. For instance, student Z expressed hesitation regarding the final result, questioning whether it is a probability value for a binary variable: “Yes, then what we finally get should not be, is a classification of 0 and 1, right? The probability of 0 is probably 40%, and the probability of 1 is 60%, right? That's probably the case, right?” [54]. However, the students demonstrated a
clear understanding of the general modeling process, enabling them to confirm and clarify related concepts, explore and select better methods guided by a clear analytical process, and ultimately develop a spiral modeling reasoning process.

3.2.2 Exploration of model selection

During the initial exploration, the students contemplated the types of predictive models that should be built. Student Y suggested “examining numerous existing models” [111]. Student X proposed the idea of “an integrated model, considering the limitations of a single model's predictive performance” [114]. Student X exhibited a thorough understanding of integrated models, explaining the concept of re-modeling a single model and providing examples to elucidate the voting and stacking methods to other students [125,127]. This initial assumption served as a guiding principle for the subsequent modeling process.

3.2.3 Understanding the data

Comprehending the meaning of the data served as a prerequisite for variable selection and data processing. The students referred to the task description document to understand the variables in the dataset, their respective representations, and the type of data [201,204,211]. Student X raised concerns about missing values, although the discussion on how to handle them remained superficial, yet still acknowledged the importance of addressing this issue [201,203]. However, student Z displayed limited ideas regarding missing value treatment, suggesting their deletion without further elaboration [202].

3.2.4 Data preprocessing

The students encountered difficulties in selecting independent variables from the real dataset, even after understanding the data's meaning. Extracting information from the dataset proved challenging due to its complex presentation. The students actively engaged in discussions on data preprocessing, particularly regarding the conversion between long and wide formats. Student X demonstrated understanding of long data, explaining that each line represents a record of behaviors for each student [295]. Similarly, student Z expressed comprehension of wide data, noting the need to combine all the students to analyze "student" as a whole [297]. The mutual conversion between long and wide format data is a crucial step in data preprocessing for data mining projects, as it facilitates subsequent analysis and modeling by providing "tidy data" (Wickham, 2014).
3.2.5 Selection of independent variables

The selection of independent variables was a crucial step in the modeling reasoning process as it bridged the understanding of data meaning and the establishment of the model, providing clues to connect the real world with the data world (Lehrer & Schauble, 2004). Through discussion, the students identified two main independent variables for predicting answering efficiency: "the answering time of each question" and "the number of questions answered within the specified time" [395]. Additionally, the students decided to initially include all variables and later choose the most influential ones through variable analysis [396]. This approach demonstrated the generation and selection of variables and attributes in the modeling reasoning process, incorporating problem definition, data understanding, and exploratory data analysis (EDA) to inform important modeling decisions.

3.2.6 Selection of sampling method

The selection of a sampling method was an indispensable step in the modeling reasoning process, as it determined the use of samples in model construction. The students not only employed "principal component analysis" and "factor analysis" as part of EDA to select independent variables [778], but also demonstrated understanding and discussion of sampling methods. Student X initially perceived sampling as "data segmentation" and described the ten-fold cross-validation method as dividing the data into ten parts, repeatedly selecting one part as the test set while using the others as the training set [780]. After extensive discussions and trials, the students ultimately chose the ten-fold cross-validation sampling method with three repetitions based on the results of model evaluation and optimization [Project report, p4].

3.2.7 Selection of modeling method

The selection of a modeling method directly influenced the modeling results. However, in educational data mining projects, specific modeling methods were often chosen based on the results of evaluation models. The students utilized data analysis software RStudio to convert complex probability formulas into specific codes, facilitating the implementation and analysis of different modeling techniques. The selection of modeling methods followed the principles of EDA, saving time and allowing the students to focus more on the data itself rather than the application of intricate statistical probability formulas. The students used RStudio and focused on understanding how the code was implemented when choosing the modeling method, considering the previously mentioned integrated modeling approach, specifically the voting and stacking methods [1145, 1151]. The choice of modeling method primarily considered the kappa value, which aligned with the EDA approach of focusing on the results and outcomes rather than becoming overly fixated on the modeling method itself [1149].

3.2.8 Model evaluation

Model evaluation closely followed the model selection process and significantly impacted all previous decisions, including the choice of independent variables, sampling method, and modeling method. The evaluation of the model focused on two key metrics: kappa and AUC [Project report, p4]. During the evaluation, the students primarily paid attention to the kappa value, while AUC was not considered at this stage. Different modeling methods produced varying kappa values, allowing for a direct comparison of their strengths and weaknesses. For instance, "regression is 0.15, decision tree is 0.17, and random forest is also 0.17" [1724]. This comparison revealed that both random forest and decision tree methods performed similarly and better than the regression method.

3.2.9 Model Optimization
Model optimization involved adjusting the input model data and parameters based on the evaluation results. The evaluation outcomes influenced the selection of independent variables and data preprocessing. The students attempted to optimize the model by adjusting its parameters, referred to as parameter adjustment. The presence of underfitting or overfitting was an important consideration during model optimization and closely tied to the model's parameters. Underfitting occurred when the model failed to predict results accurately, while overfitting referred to the model performing exceptionally well on sampled data but lacking generalization ability to apply to the entire population. The AUC value partly reflected the model's generalization ability. The students demonstrated an understanding of underfitting, overfitting, and generalization ability, engaging in discussions about these concepts: "It's under-fitting, so it lacks an ability to identify different independent variables, so he finally gets the same probability" [3076]; "Does underfitting mean no generalization ability? " [3080]; "Sure, it's too precise " [3081].

4. Discussion

Data literacy has emerged as a crucial component of 21st-century skills in the era of big data, encompassing various aspects such as modeling reasoning, data aggregation, context understanding, and data visualization (Rubin, 2019). Among these, modeling reasoning plays a significant role in establishing connections between the real world and the data world, enabling individuals to define, explain, and predict phenomena based on data and theories (Aridor & Ben-Zvi, 2017). However, there is limited understanding of how university students manifest modeling reasoning, which can impede the development of data science subjects in universities and hinder the cultivation of students' data literacy. Therefore, this study aimed to explore university students' manifestation of modeling reasoning through a case study approach.

The findings of this study reveal that the process of modeling reasoning among university students follows a spiral development form in educational data mining projects. This spiral form is observed due to the iterative and continuous optimization nature of modeling reasoning itself, as well as its role as the framework for data mining activities. The analysis process made by the students (see Figure 2) is divided into three parts: data exploration and preprocessing, modeling and application, results and feedback, which demonstrates their clear understanding of the general modeling process that enables the students to develop a spiral progress of modeling reasoning (see Figure 1). This spiral development form shares similarities with the integrated modeling approach (Bakker & Hoffmann, 2005), which also exhibits a spiral rise in its entirety. However, in the current study, university students utilized the idea of integration to link the real world and the data world at the initial stage of the modeling reasoning process, specifically during the task understanding phase. Consequently, the spiral progressive form was manifested as circular steps of modeling reasoning, accompanied by a deepening understanding of relevant concepts. Establishing the connection between data and probability is crucial in data mining activities (Konold & Kazak, 2008), which is also evident in this study through repeated understanding of sampling methods and repeated selection of modeling methods.

Teaching data science or big data remains an area with limited research (Saltz & Heckman, 2015). This study not only sheds light on the manifestation of university students' modeling reasoning but also provides insights into cultivating data literacy in universities. Project-based learning has been recognized as an effective teaching method in statistical education, promoting the development of statistical reasoning and data literacy (Ben-Zvi et al., 2018). Additionally, in the era of big data, exposing students to complex and real scientific datasets facilitates the development of statistical thinking and data literacy (Saltz & Heckman, 2015). The present study aligns with these findings by employing a project-based learning approach, providing students with real, unprocessed data to tackle authentic problems. This approach significantly enhances and develops students' modeling reasoning abilities (Konold & Kazak, 2008). Particularly in the establishment of data science programs in universities, the training of students' data literacy holds significant importance. While well-
structured problems and small dataset cases have proven effective in training students' data literacy in the past (Ben-Zvi et al., 2018; Koparan & Güven, 2015), the era of big data necessitates consideration of ill-structured problems and real, complex, and massive datasets within a project context when implementing project-based learning to cultivate university students' data literacy.

In this study, the utilization of RStudio, a data analysis software, allows students to explore the data analysis process through exploratory data analysis (EDA), transcending the limitations of understanding statistical formulas and probability models alone (Moore, 1997). While this approach enables students to focus on the analysis process in real data mining projects, it is important to investigate whether the excessive focus on result-oriented EDA may have adverse effects on the cultivation of students’ data literacy. Furthermore, computational thinking holds significant importance in modeling reasoning activities. In the educational data mining project examined in this study, the field of data science encompasses not only knowledge of mathematical statistics and specific domain knowledge (e.g., education-related knowledge), but also knowledge of computer programming (Rosenberg et al., 2019). Given the reliance on data analysis software RStudio, understanding and proficiently utilizing coding is crucial during the programming process. The students encountered difficulties and had to halt the modeling reasoning process due to unfamiliar code, emphasizing the need for computational thinking. This study highlights the importance of preparing the software environment before initiating the modeling reasoning process among university students. Consequently, further research is warranted to explore the relationship between computational thinking and modeling reasoning.

5. Conclusion

In conclusion, this study contributes to the understanding of university students' manifestation of modeling reasoning, shedding light on the cultivation of data literacy in higher education. The findings reveal that the process of modeling reasoning among university students follows a spiral development form in educational data mining projects. Besides, it is suggested to apply project-based learning and the utilization of real, complex, and massive datasets in developing students' statistical thinking and data literacy. However, the potential impacts of result-oriented EDA on students' data literacy and the significance of computational thinking in modeling reasoning warrant further investigation. By addressing these areas, educators can enhance their pedagogical approaches to effectively cultivate students' data literacy in the era of big data.

Acknowledgements

This essay was funded by the Special Fund for Postgraduates of East China Normal University to Participate in International Conferences. I would like to express my sincere gratitude to Associate Professor Bian Wu for his meticulous guidance and support throughout this research project.

References


Mathematic Learning-by-Teaching: Video Creation and Cross-Schools Staging

Te-Yang CHOU\textsuperscript{a}, Yen-Cheng YEH\textsuperscript{b} & Hong-Ming TU\textsuperscript{c}
\textsuperscript{abcGraduate Institute of Network Technology, National University, Taiwan
\textsuperscript{a}aaronzhou939@gmail.com

Abstract: With the advancement of technology and the transformation of media, traditional paper-based teaching is no longer the only way for students to learn. With streaming media becoming more and more popular, the "video" is a supplementary resource to be integrated into teaching. In this study, we based on the Interest-Driven Creator Theory and integrated video creation and the learning-by-teaching model to design learning activities. By creating videos, students can play the role of teachers to explain their mathematics concepts to their peers (Peer Tutoring) and enhance their interest in mathematics. Moreover, Students also can discuss with other peers and enhance their video. Students can construct their mathematical concepts and deepen their impressions in this activity.

Keywords: Video Creation, Peer Tutoring, Learning-By-Teaching, Cross-Schools Staging

1. Introduction

In traditional math classes, the teaching tools used by teachers primarily consist of textbooks and blackboards. The knowledge is usually imparted through direct instruction, with the teacher taking the main role in the teaching process, and students primarily absorbing the content presented by the teacher. In mathematics, units and concepts build upon each other. If the foundation is weak, students may struggle with more complex topics in higher grades, leading to misunderstandings and difficulties in connecting concepts (Lewin, 2006). This can create a bottleneck in their learning, causing a sense of aversion, fear, and anxiety towards math. Eventually, it may impact their motivation, and some might even give up on the subject. In learning mathematics, students actively construct logical concepts and deepen their understanding, addressing issues in traditional education. Practical application and hands-on experiences can cultivate learners' knowledge application, reinforcing existing knowledge (Dewey, 1944). The "learning by teaching" approach transforms student and teacher roles, enhancing abilities through instruction. Video production facilitates interactive learning.

2. Literature Review

2.1. Interest-Driven Creator Theory

The 21st century is an era of rapid change. In order to survive in such an environment, it is necessary to possess skills such as problem-solving abilities, collaboration and communication skills, critical thinking, adaptability, creativity, and imagination. Lifelong interest-driven learning is also essential (Chan, 2013; Griffin et al., 2012). To address the challenges faced in the field of education and bring about sustained impact in Asian education, Chan and a group of Asian scholars proposed the "Interest-Driven Creator Theory" (IDC) in 2018 (Chan et al., 2018). The theory consists of three core concepts: Interest, Creation, and Habit. Interest is a crucial concept that runs throughout the theory, emphasizing that learning should be driven by interests. When students are interested in their learning, it becomes enjoyable and efficient, especially when they acquire knowledge that aligns with their interests. Creation makes learning more effective and rewarding. Habit serves as an indicator of students becoming interest-driven creators because their behavioral patterns shape the kind
of individuals they become. Therefore, if students can develop a habit of creating based on their interests, they will ultimately become lifelong interest-driven creators.

2.2. Learning-by-Teaching

Learning By Teaching is an activity of "mathematical creation and sharing" that helps students reduce math anxiety. It adheres to the belief that students "can learn fast, learn slowly, but cannot fail to learn," aiming to enhance students' interest in math learning, self-confidence, and self-efficacy, thereby improving their math learning performance. Students engage in self-directed learning and grasp mathematical thinking, surpassing the traditional approach of understanding and applying conceptual problem-solving in math learning. In other words, after learning mathematical concepts, students reconstruct and explain these concepts, mastering mathematical thinking through practice. Learning By Teaching can be viewed as a form of peer tutoring, where learners help each other and learn through the process of teaching (Seel, 2011). Learners simultaneously take on the role of both learner and teacher, assisting peers as teachers. This approach effectively promotes the learning of both oneself and peers (Kobayashi, 2022). Utilizing such teaching methods reduces student pressure and encourages active participation. Peers can grow together, and explaining teaching content to others enhances one's own learning concepts (Fiorella & Mayer, 2013). Through the "Learning by Teaching" model, students can demonstrate their acquired mathematical knowledge, gain a sense of achievement and ownership, and enhance their self-efficacy and interest in learning.

3. Procedure

Video creation constitutes the core activity of this research. The filming approach involves utilizing physical tools such as paper, pens, whiteboards, and blackboards for explaining mathematical concepts and showcasing the processes of solving problems. Subsequently, recording equipment such as cameras, smartphones, and tablets with recording capabilities is employed to capture the video content. Following the completion of filming, the videos are edited to cater to students' specific requirements, utilizing technological tools. Upon the conclusion of the editing process, the works are considered finalized.

4. Research Method

Participant

The subjects of this research were fourth and fifth-grade students from three elementary schools in Taoyuan City. Specifically, there were 21 students from Class 1 of the fourth grade and 9 students from Class 1 of the fifth grade in School A, as well as 26 students from Class 1 of the fifth grade in School B, making a total of 56 students participating. Initially, the plan was for the research subjects to be fully engaged in all stages of the experiment. However, due to the impact of the pandemic, the experimental approach was adjusted to remote learning. As a result, 4 students were unable to complete the questionnaire within the designated time frame, leading to an adjusted sample size of 52.

5. Result

The analysis results of using paired sample t-test on the MSLQ motivation questionnaire for the entire student population indicate a significant difference in students' overall motivation towards mathematics after engaging in the activity of filming math lessons. It is evident that students have shown a significant improvement in their motivation for learning mathematics. In terms of the intrinsic goal orientation, the overall students' scores indicate a trend towards significance, suggesting that students are more likely to engage in learning activities due to their curiosity and desire for challenges. There is a significant difference observed in the extrinsic goal orientation aspect. (Table 1)
After the activities during the school phase, there is a significant difference in student ratings, indicating that for peers, the second video shows an improvement in video quality compared to the first video. In terms of teacher ratings, there is no significant difference after the activities during the school phase, but there is a slight improvement observed in the mean scores. However, in terms of the total score, there is a significant improvement after the activities during the school phase. Based on the above results, it can be inferred that students, after the first individual video recording and peer assessment, as well as teacher comments during the school phase, have shown a significant improvement in video quality in the second video based on the feedback received from classmates and teachers. (Table 2)

### Table 1 MSQL Motivation Questionnaire Results

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>AVG</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic goal orientation</td>
<td>52</td>
<td>3.47</td>
<td>0.837</td>
<td>-2.204</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.63</td>
<td>0.803</td>
<td>-1.235</td>
</tr>
<tr>
<td>Effort</td>
<td>52</td>
<td>3.19</td>
<td>0.740</td>
<td>-2.586**</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.42</td>
<td>0.769</td>
<td>-3.149***</td>
</tr>
<tr>
<td>Task value</td>
<td>52</td>
<td>3.13</td>
<td>0.822</td>
<td>-1.363**</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.36</td>
<td>0.901</td>
<td>-1.588***</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>52</td>
<td>3.17</td>
<td>0.825</td>
<td>-3.003***</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.27</td>
<td>0.767</td>
<td>-1.889***</td>
</tr>
<tr>
<td>Learning anxiety</td>
<td>52</td>
<td>3.11</td>
<td>0.933</td>
<td>-1.117**</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.20</td>
<td>0.912</td>
<td>-1.428**</td>
</tr>
<tr>
<td>Mathematics motivation</td>
<td>52</td>
<td>3.29</td>
<td>0.627</td>
<td>-2.413**</td>
</tr>
<tr>
<td>Post</td>
<td>52</td>
<td>3.45</td>
<td>0.566</td>
<td>-3.004***</td>
</tr>
</tbody>
</table>

### Table 2 Video Quality Sample T-test

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>AVG</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>52</td>
<td>18.76</td>
<td>2.550</td>
<td>-2.390*</td>
<td>.021</td>
</tr>
<tr>
<td>Second</td>
<td>52</td>
<td>19.60</td>
<td>2.129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>52</td>
<td>20.35</td>
<td>2.862</td>
<td>-1.824</td>
<td>.074</td>
</tr>
<tr>
<td>Second</td>
<td>52</td>
<td>20.86</td>
<td>2.275</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>52</td>
<td>39.11</td>
<td>4.550</td>
<td>-2.182**</td>
<td>.007</td>
</tr>
<tr>
<td>Second</td>
<td>52</td>
<td>40.45</td>
<td>3.432</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 6. Conclusion

Based on the aforementioned discussion, it was found that the mathematics teaching through video creation activity had a positive impact on students' interest and motivation overall. There was also an improvement in video quality. Students gained a clearer understanding of mathematical concepts and a sense of achievement during the activity. Based on the actual implementation of the activity and the results of the interviews, it can be concluded that the mathematics teaching through video creation activity is feasible in different contexts. However, there are some minor details in activity and curriculum design that can be improved or given more attention to.

### Acknowledgements

This study was funded by the National Science and Technology Council of Taiwan (NSTC 112-2423-H-008-003) and by Research Center for Science and Technology for Learning, National Central University, Taiwan.

### References


The role of individual preparation for knowledge construction in collaborative argumentation: An Epistemic Network Analysis

Wenli Chen\textsuperscript{a*}, Junzhu Su\textsuperscript{a}, Qianru Lyu\textsuperscript{a}, Siew Cheng Aileen Chai\textsuperscript{a}, Xinyi Li\textsuperscript{b}, Guo Su & Eng Ng\textsuperscript{a}

\textsuperscript{a} Nanyang Technological University, Singapore
\textsuperscript{b} Beijing Normal University
\textsuperscript{*}wenli.chen@nie.edu.sg

Abstract: Through collaborative argumentation, students gain in-depth understanding of learning content when they build on one another’s knowledge. Although individual preparation (IP) is found to be effective to foster collaborative learning, the mechanism of how IP influence the knowledge construction behavior is underexplored. This study investigated how IP influenced secondary school students in relation to knowledge construction behavioral patterns when participating in online collaborative argumentation activities. 20 students participated in two computer-supported collaborative argumentation lessons with one group with IP, and the other group without. Screen video recordings of students constructing arguments in groups during two lessons were collected and analyzed. Epistemic Network Analysis was conducted to examine students’ knowledge construction behaviors in the two lessons with and without IP. The results show that there were significant impact on students’ knowledge construction characteristics between the two lessons. Students who did not go through the IP phase tended to exhibit behaviors related to ideas refinement more than the students who went through the IP phase. The implications of how to design and implement effective knowledge construction are discussed.

Keywords: Collaborative argumentation, individual preparation, knowledge construction behaviors, Epistemic Network Analysis

1. Introduction

Learning and working environments require people to solve problems collaboratively (Graesser et al., 2018). To promote students’ collaboration competency, collaborative learning is widely used in various classrooms (Nokes-Malach et al., 2015). Studies showed that collaborative learning does not spontaneously bring benefits to students (Menekse & Chi, 2019). Various strategies have been used to promote students’ effective collaborative learning. One of the strategies is individual preparation (IP) for collaboration, which provides students with time to process the learning materials individually before collaborative learning was implemented (Tsovaltzi et al., 2015). To examine how IP influences the students’ knowledge construction behaviors, this paper analyzes students’ on-screen knowledge construction behaviors in collaborative argumentation environments with and without IP.

2. Literature Review

2.1 Computer-supported Collaborative Argumentation
Argumentation is crucial in solving ill-structured problems in the real world, which typically call for the collecting of observational data, adherence to formal logic norms, and the reasonable settling of conflicting viewpoints in discussions (Jonassen & Kim, 2010). As one of the promising approaches to improve students’ argumentation, computer-supported collaborative argumentation (CSCA) supports the sharing, constructing, and representing of arguments in multiple formats. Various online systems with various learning affordances including graph-based argumentation (Scheuer et al., 2014), representational guidance tools (Hsu et al., 2015), and micro-scripting or macro-scripting (Noroozi & Hatami, 2018) have been designed to facilitate students' collaborative argumentation processes. Graph-based CSCA was found to improve students’ learning outcomes (Chen et al., 2021). In graph-based argumentation, students use nodes or bubbles to represent different argument parts and use links or arrows to show how these parts relate to one another.

2.2 Individual preparation
In collaborative learning activities, various collaboration scripts have been designed and implemented to foster the quality of students’ contributions and improve the collaboration process (Weinberger & Fischer, 2006). As one of the effective collaboration scripts, individual preparation (IP) before collaboration is defined as “providing students with time to perform activities directed at processing the instructional material on their own before the collaboration” (Mende et al., 2021). During IP, students can prepare for the subsequent discussion, e.g., recall their prior knowledge and experience, create their own arguments, and prepare their individual ideas which they may compare or combine during the subsequent collaboration (Asterhan & Schwarz, 2007). Understanding how IP influences students’ subsequent knowledge-construction behaviors during the collaborative learning phase is of great importance.

2.3 Knowledge construction
Knowledge construction emphasizes that students construct new knowledge from social interactions (Leach & Scott, 2003). The quality of knowledge construction is dependent on how groups of students negotiate meaning, come to an agreement, coordinate tasks, and monitor the knowledge construction or group work, which are all highly correlated with group performance and project quality (Lin et al., 2016). The analysis of behavior counts across several categories has been the main focus of research on online collaborative activities (Christy & Fox, 2014). There is a need to capture the process of online collaborative argumentation and to reveal the relationships between different knowledge construction behavior.

This study investigated on students’ knowledge construction behaviors in a CSCA context to gain more CSCA design insights. The research question is: What is the difference between knowledge construction behaviors in the CSCA activities with and without individual preparation conditions?

3. Method
The case study method was applied to examine the process of one class’s collaborative argumentation that took place in two English language lessons. In the first lesson, students were asked to do a 5-minute IP followed by a 15-minute collaboration. In the second lesson, students were asked to have a 20-minute collaboration without IP. The teacher had rich experience in facilitating collaborative activities in classrooms. A CSCA environment entitled AppleTree System (Chen et al., 2021) was used to support the groups’ collaborative content editing, group management, and activity monitoring with learning analytics. When students co-constructed and refined the arguments on the Appletree system, the on-screen behaviors on the Appletree System were recorded for further analysis.

3.1 Participants
20 secondary school Singaporean students participated in this study. All the students were female students in Grade 9, aged between 14 to 15. They had experience in collaborative discussions on the CSCA system, AppleTree system. Students were randomly assigned into groups of four or five by the teacher. The students in the class are familiar with each other and have previously participated in group learning activities, although not in the same group settings as in this study.

3.2 CSCA system

The AppleTree system allows students to externalize their knowledge construction processes in an argumentation graph structure (Chen et al., 2021). On the system, students could develop graph-based argumentation to represent argument elements, and relationships between them, in which different bubbles represent ideas, claims, and evidence. Learning analytics, including social network analysis and contribution count, demonstrate on the system synchronously.

3.3 Design and Procedure

Two collaborative argumentation activities conducted across two weeks were co-designed by the teacher from the school and the researchers in this study. The students were engaged on two topics: “Foreigners are not welcomed in Singapore. Do you agree?” and “The Singapore government’s efforts have been effective in managing racial and religious tensions. Do you agree?” Each group discussed and co-constructed their group argumentation in the joint working space on the AppleTree platform for 20 minutes. Students sat together with their group members and worked on the Applee system with their personalized learning devices. In the lesson with IP, students were asked to write individual ideas for 5 minutes before they continued to write group ideas collaboratively for another 15 minutes. In the lesson without IP, students were asked to work collaboratively for 20 minutes without IP.

3.4 Data Collection and Analysis Method

To explore students’ online knowledge construction behaviors, this study collected each student’s online collaboration process data. To answer the research questions, screen recordings of all the students (N = 20) of the two lessons were collected and analyzed. Each student's screen recording was loaded into the Datavyu (release 1.3; Datavyu Team, 2014), a video coding tool to analyze students’ online behaviors. The on-screen behaviors demonstrated how students contributed and shared their ideas to the group through their behaviors.

A coding scheme was developed to examine the characteristics of the students’ knowledge construction behaviors. The unit of analysis is each action presented by the student in the online platform, such as writing one piece of evidence and monitoring the learning analytics of the platform. Students’ knowledge construction behaviors were coded based on the coding schemes adapted from the schemes of Curtis and Lawson (2001) and Popov et al. (2019) to identify different behavioral categories of knowledge construction. Seven main categories of online collaborative behaviors (Table 1) were identified: organizing, contributing, seeking input, monitoring learning analytics, refining and revising, social interaction, and activity-related individual behavior. To adapt to this study context, some codes were added to represent how students add ideas to the collaborative argumentation diagrams. In total, there are sixteen subcategories of behaviors, which are illustrated in the coding scheme in Table 1.

<table>
<thead>
<tr>
<th>Behavior categories</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizing</td>
<td>OGM</td>
<td>Organize group argumentation graph</td>
</tr>
</tbody>
</table>
Two coders coded the online screen recording data independently using the same coding scheme. Cohen’s Kappa coefficient was used to calculate the inter-coder reliability and the reliability coefficient value between the two raters was 0.704, which suggested a reasonable level of agreement between the two coders. The students’ online behavior data was organized according to the ENA data standard. The encoded data was analyzed using the ENA web application. The codes in the scheme for classifying knowledge construction behavior were chosen as the codes. Utterances in both conditions were examined for the co-occurrence of knowledge construction codes, and the relevant networks were displayed in an ENA space.

4. Results

The ENA results reveal the connections among the students’ different collaboration behaviors in different conditions for both lessons. Each node in the network graphs (Figure 1) stands for a different collaboration behavior code. The links between the various nodes show that the two codes co-occur. The red edges in Figure 2 indicate the network of the lesson with IP and show the connections between the nodes that were stronger among Internet information read (INRR), Organize individual argumentation graph (OIM), Organize Group argumentation graph (OGM), Thinking, drafting, idling (TH), Read Group members’ work (RGW), and Read one’s own work (ROWW) than the connections among other behaviors. The blue edges in Figure 1 indicate the network of the lesson without IP and show the connections between the nodes that were connected among Internet information read (INRR), Organize individual argumentation graph (OIM), Organize Group argumentation graph (OGM), Thinking, drafting, idling (TH), Read Group members’ work (RGW), and Read one’s own work (ROWW), monitoring social network analysis (MSNA), and Revise by elaboration (RE) than the connections among other behaviors.
5. Discussion and Conclusion

In summary, students in the CSCA lesson with IP mainly differ from those in the lesson without IP in which students have more behaviors connected with organizing group argumentation graph, organizing individual argumentation graph, but connected less with refining the argumentation graphs and monitoring social network analysis during collaborative activities. These behaviors represented how students’ knowledge construction happens and how these behaviors interact with each other. In the future design and implementation of argumentation activities, educators may provide specific guidance, such as encouraging more refining activities when students start collaborative without IP.

This study also examined the students’ knowledge construction behaviors of collaborative argumentation in two lessons, with and without IP. The ENA results contribute to and extend our previous IP research in argumentation activities by offering empirical evidence for supported evidence of the knowledge construction process following IP. The results confirmed that (1) students’ knowledge construction behaviors presented different characteristics with and without IP conditions, (2) students tend to have more behaviors connected to refining ideas behavior without IP conditions than with IP conditions, which corroborates with the past findings that IP may lead to the solidification of prior misconceptions (Judele et al., 2014). Besides, the different connections with refining behavior in two conditions indicates that students tend to make decisions without further refining when they write arguments in groups, which resonates that students with IP may rush to decision-making without integrating each other’s ideas (Lyu et al., 2022). In addition, though the preparatory mechanism confirms the key roles of cognitive preparation activities before the collaboration (Lam & Kapur, 2018), however, when students spend them developing their ideas during IP, it might solidify students’ cognition and that might result in a reduction of students’ refining or integrating ideas in the subsequent collaborative argumentation activities. Thus, this may result in students losing the opportunity to improving in the multiple-perspective dimension (Stapleton & Wu, 2015) despite being in a collaborative learning context, which is one indicator of an effective argumentation. There are a few limitations of this study. Firstly, the time allocated to students to work on their group argumentation artifacts is relatively short. Students had less than 20 minutes to construct knowledge and co-create the...
argumentation graph. Secondly, the sample size of the two lessons was small. Future work may extend this exploratory application of ENA to more students.

References


Students know more than they can tell: Understanding learners’ ideas of heat transfer via model revision activities

Rajashri PRIYADARSHINI*, Chandan DASGUPTA, Sahana MURTHY
IDP in Educational Technology, Indian Institute of Technology Bombay, India
*rajashri13@iitb.ac.in

Abstract: Students’ preconceptions significantly influence learning as they bring prior experiences and ideas alongside formal knowledge. Considering only explicit or conscious conceptual models can misinform teachers’ efforts. Framing the problem using both conscious (verbal-symbolic knowledge, conscious models) and unconscious (implicit models, core-intuitions) conceptual resources offer a powerful framework. This study interprets the ideas of heat transfer of two grade 11 students in depth (from a total N=9), through this framework. We look at how Nash and Payal use different conceptual resources to build an explanatory model of heat transfer in a cup through the model revision activities. We found different ideas of heat (heat as a substance, steam as unique property of heat) to be unconsciously influencing how students understand the concept. Harnessing core-intuitions into conscious models enabled Nash to build better explanatory models, whereas Payal extensively relied on her verbal-symbolic knowledge. In the macroscopic activities, both students tended to use their judgments from their sensory perceptions, going against their verbal-symbolic knowledge. Using the metacognitive prompts on Knowledge Forum also supported student’s critical analysis about their ideas before uploading them on the forum.

Keywords: Conscious models, unconscious models, intuitive knowledge, representational levels, perceptions, Knowledge Forum, explanatory models

1. Introduction

Students’ preconceptions have a serious impact on their learning of new material. They bring with them their experiences and ideas of the world while also consciously reasoning about concepts taught as formal knowledge (Piaget, 2013). Hence, considering student responses to draw only from their verbal explanations or simply identifying non-normative notions to be the result of incorrect interpretation of instruction may misinform an instructor’s efforts in harnessing students’ ‘own’ ideas (Taber,2010). Looking at the problem through the models of knowledge classified as explicit/conscious (decisions we consciously and deliberately take) and implicit/unconscious (decisions taken intuitively, mostly drawn from our mechanistic interaction with the world) can be a powerful framework to achieve this (Brock, 2015; Taber,2014). There is much evidence of learning difficulties in science education because of the ready activation of different implicit knowledge elements, overpowering the need for canonical or sophistication explanation (Brown and Hammer 2008; Taber and Garcia-Franco, 2010). Students use of conscious and unconscious conceptual resources have been reported previously. Students express ideas where unstated implicit models and intuitions may underly their conscious models (diSessa,1988; Vosniadou and Brewer, 1992; Watts & Taber, 1996). Conscious conceptual resources include the formal knowledge that students state using verbal-symbolic knowledge or use conscious models like gesturing or representing unseen elements in diagrams. Unconscious conceptual resources include core-intuitions about domain-general ideas about how things work or the implicit models like domain-specific tacit assumptions (Brown, 1993). Students created sophisticated explanatory models of magnetism when involved in unconscious ways of thinking (implicit models and core intuitions) (Brown & Cheng, 2010). Different instances such as use of verbal-symbolic knowledge abstractly, connecting verbal-symbolic knowledge only to implicit models, over reliance on
implicit models, incorrect use of core intuition or anthropomorphic models improperly impeded students’ explanatory model building processes (Brown, 2017). Implicit knowledge elements in chemical contexts (“materials naturally react”) were more about the fundamental nature of the world (Taber, 2010) and less about the intuitive sense of the mechanism as reported by diSessa’s in the context of physics (diSessa, 1993).

Despite knowing the value of looking at students’ conceptions through the implicit knowledge elements views, there is dearth of such research in science education. There are still many unanswered questions such as - do students’ intuitions align with normative models of science, what activates the use of certain intuitions over symbolic knowledge, etc. (Brock, 2015). We situate the current paper’s work across the multiple levels of representation to uncover students’ understanding of a transport phenomenon, heat transfer. We use Knowledge Forum (KF) to help us support and elicit students’ scientific knowledge building process (Lin & Chan, 2018). This paper focuses on addressing a few of the reported gaps through the research question : How do novice students build and refine their explanatory models of a scientific phenomenon on Knowledge Forum using explicit and implicit cognitive resources?

2. Methodology

Students who had completed the topic of thermal properties of matter and, heat transfer were considered for this study using a purposive sampling. A total of nine grade 11 science students, from schools in the closest cities were part of the study. The schools follow the SSC and ICSE curriculum. For the purpose of this study, we have reported the analysis of two students whose data was pertinent to our research. The study was conducted in, a studio-style collaborative classroom at IIT Bombay, India. Each group had access to a laptop, materials, and papers. Students’ discourse in the group were captured using a voice recorder, camcorder, and a screen recorder. Students’ interactions in the simulation platform, Knowledge Forum was captured by recording the screen in the laptop. Explanatory models on the Knowledge Forum and the paper sketches were used for artifact evaluation.

To elicit students’ reasoning, we used an explanatory model-based approach, where students could organically discuss and bring out their individual ideas as they worked together. The study was designed around the multiple representational levels of chemistry-the macroscopic level (experimentation with cups), microscopic level (simulation and molecular explanation) and symbolic level (explanatory models). The groups were asked to create an explanatory model on Knowledge Forum (KF), to explain the heat transfer phenomenon in a coffee cup. Before the activity began, the researchers demonstrated the different prompts and features of Knowledge Forum, and how to go about using them. The default prompts in the KF were used throughout the activities as is as they were suitable to address our study. We particularly did not direct students to use the scaffolds and they had the autonomy to use them. The explanatory models were revised on the Forum by the groups after each activity. A total duration of approximately 3.5 hours was utilized in the execution of the activities. The description of the activities is detailed in Table 1.

Table 1. Summary of Activities

<table>
<thead>
<tr>
<th>Activities</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Knowledge</td>
<td>Researcher conducts active recall about concepts of heat, temperature, and examples</td>
</tr>
<tr>
<td>Observe and discuss (single cup)</td>
<td>Researcher pours hot water and adds coffee powder. Each group is asked to observe the cup and discuss their ideas of the cup is cooling down.</td>
</tr>
<tr>
<td>Explanatory model on KF</td>
<td>Groups create an explanatory model to explain the heat transfer in a coffee cup on KF</td>
</tr>
<tr>
<td>Observe, discuss, choose (multiple cups)</td>
<td>Groups are given different cups, aluminum foil, milk, coffee, hot water to play around and choose the cup which will cool down the fastest.</td>
</tr>
<tr>
<td>Explanatory model on KF</td>
<td>Groups update their model on KF with new insights or observations</td>
</tr>
</tbody>
</table>
Molecular simulation and representation of cup
Groups are asked to create (draw/describe) a molecular representation of the chosen cup. The groups are asked to interact with the molecular workbench simulation on heat and temperature during the activity.

Explanatory model on KF
Groups are asked to update the molecular representation in the model on KF

Expertise level
Groups are asked to divide themselves each into one type of heat transfer (conduction, convection, and radiation). Each expert group is given a mix of examples of conduction, convection and radiation and asked to choose three examples of their respective expert group and justify their choice

Explanatory model on KF
Groups create an expertise model on KF

Final Discussion
All members return to their home groups and discuss their findings

Explanatory model (updating)
Final updating of models, if any by the groups

We found the framework by Brown (1993) to be best suited to analyze our students’ conceptions. The elements of the framework consist of verbal-symbolic knowledge, conscious models, implicit models, and core intuitions. Verbal-symbolic knowledge: consciously remembered verbal principles, generalizations, or equations, conscious models: conscious imaging like gestures, drawings, verbal, or written descriptions, etc. to explain different causal entities, Implicit models: tacit assumptions about a phenomenon, core intuitions: unconscious model where students use gut-level intuitions about causal interactions of different entities. Using these categories of conceptual resources, we attempt to construct an understanding of student’s ideas about the heat transfer phenomenon. Using a comparative-cross analysis method, we looked at two cases of students from two different groups to understand the ways different knowledge elements are used. During the initial screening of the data, we found these two students to be using contrasting reasoning while explaining the heat transfer process where Nash was attempting to integrate different ideas (his own and from experimenting), and Payal resorted to formally taught ideas often. We thought it to be useful to further investigate these two cases in detail. We employed a generative interpretation method based on the verbal and non-verbal discourse by students, the diagrams, the explanatory models on the Knowledge Forum platform and the sketches and explanation on the paper (Clement, 2000).

3. Findings

Ideas about heat- Heat as a substance: Nash’s explanations during his observations with the cup helped us uncover his implicit views about heat. For example, “When we touch it, we actually take some heat energy to us” tells us about how Nash might unconsciously be considering heat as a substance. The properties of substances such transitional, locational, containable, etc. became implicitly applicable while reasoning about the phenomenon. We also observed Nash’s overreliance of an observed behavior of heat transfer, steam. Nash uses conscious models gesturing the way heat moves from cup to the table. Steam was seen as a source or container of heat. In his statement, “Vapor will be transferred, and the table will absorb the heat energy,” we see how heat is considered to be contained in the vapor, confirming how heat is viewed as a substance. Payal also demonstrates this view of heat in some instances like “when we give it heat, it becomes hot, only then it will conduct right?” where she considers heat as a substance given. Even in another instance where she mentions “hot molecules of water coming from the shower,” hot molecules are considered as substances.

Ideas about heat- A phenomenon should have something unique: Unlike Nash, Payal had another view of heat should have something ‘unique’. In her statement, “There should be something unique about heat, right? We can sense heat, yeah!,” she identified ‘steam’ as a unique property of heat. She used it to reason across the macroscopic level activities while choosing if something is hot/cold. When she had to choose the cup that cooled down the fastest, Payal relied on her reasons generated by touching the cup and ‘sensing’ the heat of
This idea about heat could also have supported Payal’s reasoning with sensory perceptions.

**Ideas about heat- Agentic views:** Students unconsciously assign different sorts of agencies to the actors in the given situation. These attribute clusters operate at a deep, unconscious level of core intuition, and are not articulated in the verbal language. In Nash’s statement, “when we touch it, we actually take some heat energy to us,” the person is the initiating agent, and the touching initiates an agentic behavior that ‘takes’ the heat from the cup (affected responder, heat taken from the cup by the initiating agent (see Figure 1). In the conversation [Nash: The part of the table in contact with the cup, absorbs the heat from the cup, Riya: (The cup) it transfers the heat, Nash: “Absorb or transfer? Absorb. Transfer means it gives.”] even when his teammate suggests ‘transfer’ in place of ‘absorb,’ Nash insists on using ‘absorb’ and adds that transfer means ‘it gives’ to indicate the initiating agent as the table, as opposed to the cup.

**Sensory perception over Formal knowledge:** During the macroscopic level activities, Nash extensively used his actions of ‘feel’ and ‘sense’ to reason or make judgments. He investigated perceptually (by holding, touching and watching the materials) as well as scientifically (by systematically measuring temperatures of different coffee cups using thermometer). Even though he recalled and stated his verbal-symbolic knowledge about metal being a good conductor of heat, experimental measurements of the cups along with perceptual ‘feel’ gathered by interacting with the materials including the cup, dominated his judgment. Payal also encountered multiple instances where her formal knowledge about metals being a good conductor of heat was questioned because of her perceptual observations where she decided the degree of hotness of the cups by sensing them. Another reason why the perceptual views of both students overpowered their formal knowledge could be because of their implicit views about how conduction works. They may be considering the metal cup to be hot because it is ‘holding’ the heat, hence no heat transfer. While they were correct in observing it, their implicit ideas about heat impeded their inferences.

**Integration of conscious and unconscious knowledge:** The most noticeable difference in Nash’s and Payal’s ways of scientific reasoning was the use of the different conceptual resources. Nash, who had a sophisticated explanatory model, largely used his intuitive knowledge, perceptual observations, and meta-conceptual awareness. He rarely relied on his verbal-symbolic knowledge, using it only to support or anchor his reasoning. We could also see him invoke conscious models to explain the mechanism of the phenomena like movement of steam or energy from the cup to the surroundings. Unlike Nash who relied on his verbal-symbolic knowledge occasionally as a support, Payal constantly relied on her verbal-symbolic knowledge not as a support but as a main means to build the explanatory model. We could rarely see her draw on her intuitive knowledge. We also see a lack of meta-conceptual awareness in Payal’s model iteration process.

**A framework of Nash’s ideas of heat transfer:** Nash’s use of multiple conceptual resources can help us understand how the resources interacted to build his understanding of heat transfer. Using his most dominant view of heat, we drew a representation to explain the interplay of all the resources at the macroscopic level activity as depicted in Figure 1. The verbal symbolic knowledge helped him create two different conscious explanatory models of heat transfer. These conscious models were implicitly dictated by his views about heat being an object(substance). Core intuitive resources such as attribute clusters underlined the explanation of how Nash saw hand causing heat transfer. This representational diagram helps us “see” briefly the way different conscious and unconscious conceptual resources could work towards building an understanding of a phenomena for a student.
4. Discussion and Conclusion

From the findings of both cases, we observe how students employ diverse conceptual resources to construct scientific ideas. Nash developed a sophisticated explanatory model of heat transfer by effectively integrating his unconscious conceptual resources into his conscious thinking. Core intuitions and implicit models played a significant role throughout his sense-making process, with verbal-symbolic knowledge mainly supporting his model creation. Nash initially leaned on verbal-symbolic knowledge, using definitions of heat and temperature to initiate his reasoning. His modeling process remained predominantly conscious, using gestures and diagrams to convey his ideas. In contrast, Payal struggled to access intuitive conceptual elements and consistently relied on verbal-symbolic knowledge. While she did revise her explanatory models, research shows that successful explanatory model creation involves the incorporation of implicit models and core intuitions. Students who solely rely on verbal-symbolic knowledge or inaccurately connect it to implicit models hinder their explanatory model development (Brown & Cheng, 2010).

In both cases, students' implicit models of heat influenced their scientific reasoning about heat transfer. Students often view heat as a substance or object, aligning with early caloric theory. Previous research on high school students' misconceptions has also identified this tendency to conceptualize heat and temperature as a substance (Erickson, 1985). Payal searched for a unique aspect in heat, almost treating it as a distinct "sense." Nash, while not explicitly focusing on heat's uniqueness, also employed the notion of "sense" in explanations. These notions might hinder students' comprehension of mechanisms like heat transfer. Instead of seeking common underlying mechanisms, students become fixated on visible behaviours when reasoning (Wagner, 2006; Taber, 2010). Instructors can seize such instances to illustrate mechanistic similarities across phenomena. Students' heavy reliance on perceptual experiences, particularly in macroscopic activities, impacted their verbal-symbolic knowledge. Because heat is a phenomenological concept, rooted in sensory perception and firsthand experience, it holds more persuasive power than formal knowledge. This tendency led students to generate alternative explanations to match their observations—a trend not observed in studies on electricity or magnetism (Brown, 2017).

During the activities, Nash systematically developed ideas from various sources, integrating them across representational levels. He connected different heat transfer types in his group's model, showcasing a sophisticated understanding of heat transfer (Sunyono et al., 2015). Nash's engagement with prompts like 'my theory' and 'I need to understand' on
Knowledge Forum encouraged metacognitive thinking, leading to in-depth inquiry and refining his explanatory model (Tan & Loong, 2005). This approach could be valuable in large classrooms where individual attention is challenging. Our study highlights how students' conceptual resources influence learning about scientific phenomena, suggesting the benefit of merging intuitive and formal knowledge. Prompting students to explain divergent outcomes for similar situations can enhance their epistemic motivation (Taber, 2010). Connecting scientific principles to everyday experiences can further improve students' epistemological consistency (Lemmer et al., 2020), with potential applications in various phenomena. This article does not explore group dynamics' impact on students' resources, limiting its findings' strength within the group activity context. The analysis acknowledges potential bias in student utterances due to the interpretative approach taken. Despite considering the context, misleading interpretations are possible. The study's scope is confined to two cases; future research will encompass additional cases for robustness.

References
Impact of 360°VR on Pre-Service Teachers’ Empathy——Taking Educational Equity as an example

Yan-Jun CHEN*, Yi-Ling HU & Bian WU
Department of Education Information Technology, East China Normal University, China
*904553828@qq.com

Abstract:

Under the goal of promoting a more equitable, inclusive, and resilient education system, how to use innovative learning technologies to develop teachers’ empathy is a hot research topic in the field of educational equity. Recent research has emphasized the potential of 360°VR to bolster empathy. The underlying mechanisms of such outcome are, however, underexplored. The present study applied three types of educational equity learning materials, including 360°VR, 2D video, and graphic, to pre-service teachers’ empathy development. Randomly assigned participants viewed a documentary featuring a group of children living in rural and urban areas in a 360°VR or 2D video or graphic format, and measures of empathy (in terms of perspective taking and empathic concern), immersion, enjoyment and discomfort were collected. Results indicated that the virtual reality experience resulted in a higher level of empathy but this effect was not significant in terms of perspective taking. Immersion is the process of linking virtual reality experience with empathy, and 360°VR brings higher immersion by triggering the illusion of spatial presence which was associated with higher levels of empathy. Implications of the study include: it is feasible to use 360°VR to develop teacher empathy training, but needs to be combined with effective instructional activity design and scaffolding design; using empathy as an entry point can develop a two-way structure for teachers’ cognitive and affective cultivation of educational equity; and teachers’ empathy cultivation needs to build a long-term, incremental cultivation system.

Keywords: Empathy, 360-degree video, education equity, pre-service teacher

1. Introduction

As pre-service reserve teachers, normal students will face the development of cross-regional and cross-cultural diversified teaching activities in the future. The students they train also come from regions where the development of educational resources is unbalanced. Their empathy ability from the perspective of educational equity has been determined as Key characteristics of being effective in diverse teaching as described above (Darling-Hammond, 2000). In this regard, empathy is considered as part of a caring, supportive and responsive teaching style (McAllister, 2002), which is beneficial for teachers to respond to students' cognitive and emotional needs, and to care about the impact of students' life circumstances and prior knowledge on their academic development, especially those in less developed areas, to achieve student-centered teaching practices, and to build trust in relationships with students. (Warren, 2013). However, research on the empathy of primary and secondary school teachers from the perspective of educational equity shows that most teachers' knowledge and practice in educational equity are not in sync. Although some teachers acknowledge that equity is one of the basic principles of interaction with students in the classroom Although some teachers acknowledge that equity is one of the basic principles of interaction with students in class, they still have strong biases in teaching practice in terms of students' gender, personality and social background.
360°VR is known as an empathy machine in terms of enhancing empathy because of its immersive, interactive nature (accessed Sept. 15, 2019). However, the current research on 360° VR technology and teacher training focuses on the training of teachers' teaching knowledge and skills in VR environment, and involves less in the field of emotional cognition such as empathy cultivation. Based on this, this study attempts to compare the effects of learning materials presented by different streaming media on the empathy ability of pre-service teachers, analyze the possible factors affecting the empathy ability of pre-service teachers. The following research questions are proposed:

RQ1: What are the differences in pre-service teachers' empathy from an educational equity perspective in the presentation of 360VR videos, 2D videos and graphic materials?

RQ2: How do user experiences such as immersion and engagement in 360VR environments impact on empathy? And what are the interactions within them?

2. Method

A total of 88 participants (46.2% identifying as female, 52.3% male, and 1.5% other) were recruited from a Normal University in China. The experimental group presented the learning materials in 360VR (360VR GROUP, n=28) and the control group presented the same learning materials as the 360VR condition in 2D format (2D GROUP, n=32) or graphic format (Graphic GROUP, n=32) respectively. The learning material is a short documentary comparing the living and educational environment of rural and urban children. Participants in the 360VR group wore VR headset and observe the video environment from inside the spherical space created by VR. The 2D format was created by cropping the center frame of a panoramic version of the 360 documentary, retaining a third of the original field of view. The graphic version was created by intercepting key frames from the 2D video with a narration adapted from the video commentary.

Participants completed a total of three questionnaires before and after the experiment. The personal information questionnaire used before the experiment recorded the participants' personal information such as gender, major and grade, which was used to record information of participants and to test the success of the randomization. The empathy scale used after the experiment is adapted from Davis' Interpersonal Reactivity Index (Davis, 1983). Two main dimensions of Perspective Taking and Empathic Concern (six items in total) were selected for measurement. The IVR user experiences questionnaire used after the experiment was based upon previously established instruments developed to measure immersion (Charlene Jennett, 2008), engagement (Jeanne H. Brockmyer · 2009) and discomfort (Sara Ventura, MD · 2020), a total of 10 items are included.

3. Results

Research question 1 inquired what are the differences in pre-service teachers' empathy from an educational equity perspective in the presentation of 360VR videos, 2D videos and graphic materials. The ANOVA test for the empathy scale shows there was a significant difference in empathy among the three groups (F=3.128, p=0.005<0.05) and this difference was also significant in the dimension of Empathy Concern (F=5.901, p=0.013<0.05), which is similar to previous studies. Moreover, no significant difference in Perspective taking among the three types of material presentation, possibly because the Perspective taking dimension is a more stable trait at the level of personal values that is not influenced by short-term experimental training. Using the games-howell method to test empathy after the event (see Table 2), the result shows that the empathy of learners using 360VR learning materials is significantly higher than that of learners using 2D materials, and is not lower than that of learners using graphic materials. The use of 360VR technology has a good effect on the cultivation of learners' empathy ability, which answers the first research question. In contrast to previous studies, the results also indicated that the graphic materials were more effective than the 2D materials in promoting empathy.
Research question 2 inquired how do user experiences such as immersion and engagement in 360VR environments impact on empathy. The result shows the correlations of the focal variables involved in the study. The results show that in terms of the internal correlation among VR user experiences, discomfort and immersion showed a statistically significant negative correlation \( (r=-0.514, p<0.01) \) and immersion and engagement showed a significant positive correlation, \( (r=0.395, p<0.05) \). In terms of the external correlation between VR user experiences and empathy, the correlation between empathy and immersion was also statistically significant, \( (r=0.387, p<0.05) \). These findings suggest that positive IVR user experiences should be helpful in promoting empathy. Conversely, as well as negative IVR user experiences might have negative impacts. There is an internal offsetting relationship between these two experiences.

4. Discussion

Virtual reality experience resulted in a higher level of empathy and also led to higher levels of empathic attention in empathy for an individual featured in the virtual reality material compared to control conditions presenting the same material in a non-virtual reality formats. This finding is similar to previous research findings (Schutte, 2017 & Chien, 2020). In virtual environment, viewers may feel their emotions or situations more strongly as they are close to them and sharing the same space. Contrary to our initial hypothesis, the virtual reality experience had no significant effect on perspective taking. According to Davis' research, the perspective taking is different from the empathic care and is associated with relatively less emotional reactivity and higher interpersonal functioning and self-esteem. The virtual reality materials selected for this study focus more on the emotional aspects of urban and rural education equity stories, rather than promoting the cognitive perspective, which is one of the reasons why the choice of perspective for learners in the 360VR environment has not been significantly improved.

Furthermore, the immersion triggered by virtual reality reinforces various emotional and cognitive responses, thus enhancing empathy. 360VR blurs the boundary between the real world and the virtual world by creating a virtual reality environment that closely resembles the real physical space, viewers gain an immersive virtual reality media experience, generating the feeling of sharing the same time and space as the characters in the video, and promoting empathy through the sharing of perspectives. The discomfort of virtual reality experiences, such as dizziness and vomiting, can inhibit viewers' empathic feelings. This is because the high-dimensional perceptual nature of virtual reality technology can place a degree of cognitive load on the learner, resulting in negative reactions such as vertigo. As a psychological trait, empathy is related to an individual's physical condition. Negative physical experience will inhibit an individual's cognitive and emotional experience, thereby preventing the emergence of empathy to a certain extent. In previous studies, the immersion provided by virtual reality environments was a key factor in triggering high levels of empathy. However, the conclusions of this study show that the advantages of virtual reality technology related to immersion do not necessarily lead to a high level of empathy. Immersion itself is not enough to generate concrete and contextual cognition, and it must be accompanied by high-quality learning content. In these teaching practices, we should not be limited to simply converting 2D video into panoramic video playback, but should take advantage of the immersion and participation of 360VR, take advantage of technology to carry out effective teaching activity design, and integrate teaching intervention into pre-service teachers' VR experience.
Enhancing student teachers’ collaborative interdisciplinary design through knowledge-building activities

Pei-Yi LIN a*

a Department of Education and Learning Technology, National Tsing-Hua University, Taiwan

*peiyilin@mx.nthu.edu.tw

Abstract: This study aims to investigate student teachers’ interdisciplinary design collaboration and integrated teacher knowledge in a knowledge-building environment. A class of 14 participants was divided into four groups to discuss interdisciplinary learning design ideas using Knowledge Forum. Data include (1) a pre- and post-questionnaire designed to investigate the participants’ interdisciplinary design capacities (e.g., teacher efficacy, epistemic fluency, interdisciplinary design communication, integrated teacher professional knowledge) and (2) records of online activities such as numbers of notes generated, modified, and read. The findings showed that student teachers’ perception of teacher efficacy to design interdisciplinary learning, epistemic understanding, interdisciplinary design communication, pedagogical content knowledge (PCK), technological pedagogical knowledge (TPK), and integrated technological, pedagogical, and content knowledge (TPACK) were significantly enhanced. The online knowledge-building activities were more likely to relate to student teachers’ epistemic understanding and integrated teacher knowledge (i.e., TCK and TPK). The significance of practice-based, design-oriented activities for developing interdisciplinary collaboration in teacher preparation is discussed.

Keywords: teacher efficacy, epistemic fluency, interdisciplinary design communication, teacher knowledge, knowledge-building activities

1. Introduction

In today’s world, interdisciplinary design practices are crucial in preparing students to tackle multifaceted, complex problems. The interdisciplinary curriculum combines skills and content from several subjects, allowing students to make meaningful connections across subjects. Traditional teacher training programs focus on single-disciplinary knowledge. Teachers should break the discipline boundary to consider students’ interests when creating interdisciplinary curricula and relate the content to their lives (Chiang et al., 2020), recognize and participate in pedagogical design discourses to enhance their epistemic understanding when approaching problems and goals with different knowledge bases (Bopardikar et al., 2021). When teachers collaborate to create lesson plans from different disciplinary knowledge, they externalize design thinking and contribute design ideas based on their expertise to communicate and develop new understanding. Research is needed to ascertain how teachers from different fields establish a shared account of design tasks to synthesize their ideas to create interdisciplinary learning designs.

Interdisciplinary learning design entails how students learn to apply, conceptualize, model, and generate solutions from multiple disciplines (Capobianco et al., 2018). Teachers must integrate various subject knowledge and pedagogies to engage students. Technologies can support student engagement and provide adaptive scaffolds to explore an in-depth understanding of what they have learned (Ouyang & Chang, 2019). Hence, teachers’ design capacities to integrate content knowledge from multiple disciplines and pedagogical uses of technologies are critical (Chai et al., 2011). TPACK is the synthesized knowledge for integrating technologies into learning content and classroom teaching. When teachers design interdisciplinary learning, they make a pedagogical decision in deciding how to present the learning content (i.e., PCK), consider how to use technologies to address specific learning contents (i.e., TCK), discuss how to use technologies to support student learning (i.e., TPK),
and knowledge of using technologies to facilitate students learn and create an understanding of subject matters (i.e., TPACK) (Chai et al., 2011). When teachers design, they share their understanding from different fields and consider how to engage students to approach solutions to the problems that may require fluent design talks about integrating technologies, pedagogies, and multiple learning content in design communities.

Knowledge building could facilitate student collaboration and take collective responsibility for improving ideas and pursuing more coherent explanations (Scardamalia & Bereiter, 2014). It enables student teachers to work adaptively with their initial design ideas, and through continuous discourse and redesign, they view community discourses as progressive problem-solving opportunities to generate better ideas and solutions (Yang et al., 2020). This study investigates student teachers’ interdisciplinary lesson design by focusing on their efficacy in designing interdisciplinary learning, epistemic fluency, and interdisciplinary communication in developing integrated technological, pedagogical, and content knowledge practices through knowledge-building activities. The main research questions of this study are: Did student teachers enhance their efficacy, epistemic fluency, interdisciplinary design communication, and teacher knowledge across knowledge-building activities? Were their online knowledge-building activities related to the above variables?

2. Methodologies

2.1 Participants and learning activities
The participants were 14 students in an interdisciplinary learning course from a national university in Taiwan. They were heterogeneous and divided into four groups, three to four people a group, specializing in different domains. In this course, students were required to design an interdisciplinary learning micro-teaching plan based on sustainable development goals (SDGs) and present the lesson plan as course reports in a seven-week-long learning. The instructor adopted knowledge-building principles to engage students in design activities. For instance, in the first week, the instructor encouraged the students to provide ideas based on the SDGs topic they were interested in in Knowledge Forum (KF). Then, students read and discussed their initial ideas with group members before deciding on their group micro-teaching topic. Each week, there were lectures and group discussions based on the weekly assigned papers for two hours; the participants recorded their design ideas in KF and discussed them for an hour. The data of this study focused on the latter part.

2.2 Data source
Data sources included (1) online activity logs in Knowledge Forum and (2) pre- and post-questionnaires, including teacher efficacy to design interdisciplinary learning, epistemic fluency, interdisciplinary design communication (variables were validated in Lin et al., 2022), PCK, TCK, TPK, and TPACK (variables were validated in Chai et al., 2011). All items were scored on a 7-point Likert scale (i.e., 1 = strongly disagree to 7 = strongly agree).

3. Results

3.1 Student teachers’ perception of interdisciplinary design capacities
A paired sample t-test was conducted to analyze student teachers’ perceptions of efficacy, epistemic fluency, interdisciplinary design communication, and the development of integrated teacher knowledge. The results indicated significant differences in teacher efficacy (pre-test: $M = 5.86, SD = 0.81$; post-test: $M = 6.64, SD = 0.44$; $t = -3.91, p < 0.01$), epistemic fluency (pre-test: $M = 5.33, SD = 1.06$; post-test: $M = 6.36, SD = 0.30$; $t = -4.14, p < 0.01$), interdisciplinary design communication (pre-test: $M = 5.62, SD = 0.93$; post-test: $M = 6.60, SD = 0.47$; $t = -4.07, p < 0.01$), PCK (pre-test: $M = 5.14, SD = 1.25$; post-test: $M = 5.98, SD = 1.00$; $t = -2.65, p < 0.05$), TPK (pre-test: $M = 5.29, SD = 1.27$; post-test: $M = 6.26, SD = 0.64$; $t = -2.83, p < 0.05$), and TPACK (pre-test: $M = 5.38, SD = 1.31$; post-test: $M = 6.23, SD = 0.67$; $t = -2.41, p < 0.05$). However, TCK (pre-test: $M = 5.39, SD = 1.29$; post-test: $M = 5.33, SD = 0.32$; $t = 0.19, p = 0.85$) showed no significant enhanced.

3.2 Relations among student teachers’ KF activities and perception of interdisciplinary
design capacities
This study calculated the changes in student teachers’ perceptions of interdisciplinary design collaboration (i.e., post-test – pre-test) and analyzed their correlations to online interdisciplinary design discussion (e.g., # of notes generated, modified, and read). The Spearman’s rho correlation analysis results indicated that idea generation was significantly related to their TCK and TPK ($\rho = 0.60$ and $\rho = 0.59$, $p < 0.05$), and idea modification was significantly related to epistemic efficacy ($\rho = 0.62$, $p < 0.05$).

4. Discussion
Previous literature on interdisciplinary curriculum and instructional design focuses on students' learning processes and competencies (e.g., Jiang et al., 2019), offers instructional guidance (e.g., Prain et al., 2023), and documents teachers' collaborative design processes (e.g., Bopardikar et al., 2021) through case studies. However, how student teachers collaborate in designing interdisciplinary learning and integrating multiple content knowledge, pedagogies, and technologies needs to be explored. The present study investigates student teachers' interdisciplinary design collaboration in online knowledge-building activities by employing pre- and post-questionnaires to assess their perceptions. The results indicated a significant enhancement in their efficacy in designing interdisciplinary learning, interdisciplinary design communication, epistemic fluency, PCK, TPK, and TPCK.

Regarding their participation in knowledge-building activities, student teachers' idea generation was related to the change in TCK and TPK. This may imply that the online knowledge-building environment may make their design ideas explicitly to discuss integrating technologies, learning content, and pedagogies. Furthermore, this study shows the potential relation of student teachers' idea modification relates to their epistemic efficacy. When they engage in collaborative interdisciplinary design, they share initial design ideas from different disciplines, and after discussion, they may formulate new information or restructure their existing knowledge. This study proves that knowledge-building activities enrich opportunities to support the social construction of interdisciplinary knowledge and learning activities.

Acknowledgments
This work was financially supported by The National Science and Technology Council, Taiwan (NSTC 111-2410-H-007-003-MY2).

References


From Individual Ideation to Group Knowledge Co-Construction: Comparison of High- and Low-performing Groups

Wenli CHEN\textsuperscript{a*}, Guo SU\textsuperscript{a*}, Xinyi LI\textsuperscript{b*}, Qianru LYU\textsuperscript{a}, Junzhu SU\textsuperscript{a}, Siew Cheng Aileen CHAI\textsuperscript{a} & Eng Eng NG\textsuperscript{a}

\textsuperscript{a}National Institute of Education, Nanyang Technological University, Singapore
\textsuperscript{b}Faculty of Education, Beijing Normal University, China
\textsuperscript{*wenli.chen@nie.edu.sg}
\textsuperscript{*nie22.sg4560@e.ntu.edu.sg}
\textsuperscript{*li_xinyi@mail.bnu.edu.cn}

Abstract: This study compares the high- and low-performing groups’ knowledge co-construction process in the context of computer-supported collaborative argumentation from epistemic, argument, and social perspectives. Product analysis, lag sequential analysis, Sankey diagram visualization, and social network analysis were used to analyze groups’ written argumentation artefacts, on-screen behaviors, and online interactions. Results show that the high-performing group students demonstrated a higher level of engagement and cognitive elaboration than the low-performing group. The high-performing group was more competent in integrating various argumentation elements than the low-performing group. And the students in the high-performing group tended to contribute equally to their group work. The implications of the findings in designing and implementing knowledge co-construction activities are discussed.

Keywords: Knowledge co-construction, collaborative argumentation, computer-supported collaborative learning

1. Introduction

Collaborative argumentation is an effective way for students to learn together. It deemphasizes the social interactions that take place when students discuss divergent claims and come to a consensus (Gao et al., 2022; Han et al., 2021). Existing research studies show that collaborative argumentation benefits students’ knowledge construction in classroom practice (Asterhan & Schwarz, 2016). Computer-supported collaborative argumentation (CSCA) supports students to practice argumentation simultaneously through writing and discussion (Baker, 1999). In CSCA, students with differing views critically debate a subject. Ideally, they come to an agreement and collaborate to construct argumentative knowledge after comparing, assessing, and synthesizing multiple points of view (Han et al., 2021). In order to learn about argumentation and content knowledge, students generate and co-construct arguments by interacting with their learning partners (Andriessen et al., 2003). Argumentative knowledge co-construction is one of the important competencies for joint problem-solving (Vuopala et al., 2019).

To examine multiple dimensions of the knowledge co-construction process in computer-supported collaborative learning (CSCL), Weinberger and Fischer (2006) proposed a framework that contains the epistemic, argument, and social-mode dimensions. The epistemic dimension describes how learners work on the knowledge construction task (Fischer et al., 2002). This dimension looks into whether students are working on the job at hand (engaged in on-task conversation) or if they are off-task. The argument dimension focuses on how learners construct arguments with regard to defining formal relationships between specific
elements of arguments, such as claims, data, and warrants (Fischer et al., 2007). The dimension of social modes indicates how learners interact with each other.

Previous studies attempted to unpack the complex and dynamic process of group students’ argumentative knowledge construction. For example, Ouyang et al. (2023) developed an integrated discourse analysis approach to examine how the group’s discourse movements, structures, and turn-taking processes evolve over time by analyzing the collaborative argumentation from cognitive, metacognitive, and social perspectives. Lu et al. (2011) explored whether discourse moves, sequences, and participant features were related to two types of justifications (evidence and explanations) during online argumentative discussions. Most of the existing studies attempted to study the group’s collaborative argumentation process itself without considering how individual students in the group contribute to the knowledge co-construction of the group. Compared with direct collaboration, adding individual ideation before collaboration may require different epistemic activities, making it worthwhile for a detailed investigation (Valero et al., 2022).

2. Literature Review

Knowledge co-construction refers to collaborative processes in which learners interact with one another in order to build new knowledge and enhance the knowledge and understanding of each individual student (Janssen et al., 2009). These productive interactional processes include asking thought-provoking questions, exchanging ideas, negotiating and providing justified arguments in order to build new knowledge (Arvaja et al., 2007; Sawyer, 2006). There is evidence that active participation in knowledge co-construction activities can result in high-quality learning outcomes (Wang, 2009). Previous research has suggested that knowledge co-construction activities can have a positive impact on both group productivity and learning outcomes (Weinberger et al., 2007).

Collaborative argumentation is a group-based exercise in which students work together to construct an integrated group view on a discussion topic (Noroozi et al., 2012). Learners benefit from collaborative argumentation because they obtain new ideas and information while comparing, assessing, and synthesizing different points of view based on evidence and rationale (Han et al., 2021). Furthermore, collaborative argumentation is helpful in the development of competencies for arguing and collaborating with others, both of which are required for working and living in the future society (Noroozi et al., 2012).

Typical collaborative learning processes commence with the group or end at the group level. In fact, participating in individual ideation before collaboration can benefit both individuals and groups. By engaging in individual ideation before intra-group knowledge co-construction, each group member’s ability to recall prior knowledge is enhanced, fostering a greater number of ideas that can subsequently contribute to the overall intra-group synergy (Chen et al., 2021). Additionally, individual ideation promotes increased levels of participation, discourages the presence of free-riders, and equalizes the level of collaboration among group members (Isohätälä et al., 2017).

This study analyzes the process of collaborative argumentation of two groups (one high-performing group and one low-performing group) from individual ideation to group knowledge co-construction from epistemic, argument, and social dimensions. To be specific, the research questions are:

- RQ1: What are the sequential patterns of epistemic knowledge co-construction behaviors of the high- and low- performing groups in CSCA?
- RQ2: How did the high- and low-performing groups’ argument structures develop?
- RQ3: What are the social interaction networks of high and low-performing groups’ argumentative knowledge co-construction?

3. Methodology

3.1 Context and Participants
The study took place in a secondary school English class in Singapore. A total of 24 secondary school (15-year-old) students (male =12, female =12) participated in this study. The teacher randomly assigned participants to 6 groups (4 members in each group). The participants in this study possess experience in collaborative learning but lack familiarity with argumentation. The experiment occurred in a blended learning environment; the teacher and students interacted offline in a physical classroom while they engaged in collaborative arguments on an online platform using their personal learning devices (PLDs). The teacher's expertise in computer-supported collaborative learning (CSCL) contributed to the experiment's instructional design. The learning objective is applying English language skills in an argumentation context and improving students' argumentation skills and critical thinking, as highlighted by 21st-century student outcomes. IRB approval has been obtained from the university of the first author before data collection commenced.

3.2 Procedure

The participants were asked to collaboratively co-construct an argumentation diagram in Appletree, an online collaborative argumentation platform (see Figure 1. [http://www.appletree.sg/] (Chen et al., 2013; Tan & Chen, 2022).

The argumentation topic is “AI will make our lives at home more enjoyable. How far do you agree?” selected from the English lesson syllabus. The whole lesson lasted for 1.5 hours which included teachers’ instruction, and students’ collaborative argumentation. The students went through a 5-phase Spiral Model of Collaborative Knowledge Improvement (SMCKI) (Chen, et al., 2021). Phases 3, 4, and 5 are inter-group feedback, intra-group refinement, and individual achievement, which touch on students’ feedback literacy and exceed argumentative knowledge. To better understand argumentative knowledge construction, this study focuses on the first two phases which lasted for 30 minutes in total: 1) Individual ideation (10 minutes): students individually generate ideas and take a position for the augmentative topic; 2) Intra-group synergy (20 minutes). Group members co-constructed an argumentation diagram by integrating all individual ideas through discussion and negotiation.

![Figure 1. Appletree working space. Phase 1 individual ideation (left), blue bubble showcases the idea generated separately by each group member; Phase 2 group synergy (right) shows the argumentation work co-constructed by all group members. The dark blue bubble is the main claim; the green/red represents for/against the relationship between the two bubbles.](image)

3.3 Data Collection and Analysis Method

Two types of data were collected: the group’s argumentation diagrams (artifacts) in AppleTree, and the computer screen recordings from each group member of the 2 groups. Multiple analytical methods were used to examine the students' knowledge co-construction. The study used product analysis (PA) (Novak & Cañas, 2008) to assess the quality of all groups’ argumentation diagrams in terms of four dimensions, i.e., clarity, multiple perspectives, selection of evidence, and depth of elaboration (Marzano et al., 2000; Stapleton & Wu, 2015; Tan & Chen, 2022). The unit of analysis is the argumentation diagram of the group. Each
dimension is rated on a scale ranging from 0 to 3. Two of the authors scored the argumentation diagrams independently (Cohen’s $K = 0.70$). Two groups with the highest and lowest quality of argumentation artefacts were selected in this study to compare the process of knowledge co-construction between the high- and low-performance groups. The high-performance group’s score was 8 (clarity = 2, perspective = 2, evidence = 2, elaboration = 2), and the low-performance group’s score was 5.5 (clarity = 1.5, perspective =1.5, evidence = 1.5, elaboration = 1).

First, the coding scheme for on-screen behaviors was adapted from Chuy et al. (2011)’s explanation-seeking dialogue framework and the Claim-Evidence-Reasoning (CER) model (Toulmin, 2003). We take one behavior as a unit, for example, student add one subclaim, then we code it as ABC. The explanation-seeking dialogue framework includes five epistemic behaviors: 1) formulating thought-proving questions; 2) theorizing; 3) obtaining evidence; 4) working with evidence; and 5) synthesizing and comparing. This framework was most used in the knowledge building context (e.g., Knowledge Forum online platform). CER is an argumentation pattern including three argumentation elements: "claim", "evidence", and "reasoning". An argument's main goal is to justify a claim with evidence and reasoning (Osborne & Patterson, 2011). The coding scheme is shown in Table 1. The coding process was iterative. The second and third authors coded the screen recordings separately and held multiple meetings to resolve discrepancies. The inter-rater reliability, as measured by kappa coefficient, between the two coders was 0.761.

| Table 1. Coding Scheme for Knowledge Co-construction in Collaboration Argumentation |
|----------------------------------|------------------|------------------|
| Category                        | Sub-category     | Code             |
| Contribute                      | 1. Add a claim or sub-claim | ABC             |
|                                 | 2. Add new facts/examples (evidence) to support a claim/sub-claim | ASE             |
|                                 | 3. Add new facts/examples (evidence) to oppose a claim/sub-claim | AOE             |
|                                 | 4. Provide elaboration to explain evidence/sub-claim | ELA             |
| Seeking input                   | 5. Seek help from discussing with other group members | DIS             |
|                                 | 6. Seek input (from internet/discussion) | INP             |
| Revising                        | 7. Copy & paste | C&P              |
|                                 | 8. Rephrase/Revising previous work | RE              |
|                                 | 9. Delete information | DE              |
|                                 | 10. Change link/position | CH              |
|                                 | 11. Synthesize available evidence/elaboration (bubble) | SYN             |
| Off task                        | 12. Discuss irrelevant things/play with the system | OFF TASK        |

4. Results

4.1 The Sequential Patterns of Argumentative Knowledge Co-construction

The results of on-screen behavioural coding consist of a totally 111 and 67 behaviors for the high- and low-performing groups, respectively. The frequency and percentage of each coded behaviour were calculated for the two groups (see Table 2). In the high-performing group, the most frequent on-screen argumentative knowledge co-construction behaviour was seeking input from the internet (INP), whereas and least frequent behaviour was off-tasking (OFFTASK). By contrast, the most frequent on-screen behaviour of the low-performing group was discussing (DIS) while behaviours like synthesizing existing and new information (SYN) and copying and pasting (C&P) were not identified.

| Table 2. Frequencies and Percentages of On-screen Argumentative Behaviors between the High- and Low-performing Groups |
|---------------------------------------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | INP | DIS | SYN | RE | C&P | ASE | ABC | OFF TASK | DE | ELA |
| High | 27 | 14 | 9 | 17 | 11 | 7 | 5 | 4 | 7 | 10 |
Lag Sequential Analysis (LSA) (Bakeman & Gottman, 1997) was conducted based on the coding results to understand the sequential pattern of students’ knowledge co-construction from the epistemic dimension. Lag sequential analysis (Bakeman & Gottman, 1997) examines the occurrence and temporal order between different learning behaviors (Kapur, 2011). The coded knowledge co-construction behaviour sequence was imported into the GSEQ 5.1. The z-score of every transition from one behaviour to another was reported in the LSA results, indicating the significance level of each behaviour sequence (z-score > 1.96, Bakeman & Gottman, 1997). A behavior transition diagram was created to visualize the significant sequences in each group.

Figure 1 visualizes the on-screen behavioral transition patterns in terms of their argumentative knowledge co-construction of the high and low-performing groups. There were 4 and 2 significant behavioral sequences identified in the high and low performing group respectively. Firstly, the predominant patterns of high performing group’s argumentation co-construction were "seeking input from the internet or discussion (INP)" → "synthesizing internet information with existing evidence or reasoning (SYN)" and "synthesizing internet information with existing evidence or reasoning (SYN)" → "seeking input from the internet (INP)". The findings show that the high-performing group students searched for information online and then integrated that information into their proposed ideas. They tended to seek information from the internet after synthesizing ideas. Another significant transition pattern for the high-performing group was "seeking input from the internet (INP)" → "adding new evidence to support a claim or sub-claim (ASE)". The high-performing group raised evidence for claims after searching for relevant information. On the contrary, the low-performing group directly proposed claims and sub-claims after seeking information from the Internet without deeper information processing according to the transition pattern "seeking input from the Internet (INP)" → "adding a claim or a sub-claim (ABC)".

In addition, high-performing group students continued to construct knowledge via rephrasing or revising previous ideas after copying and pasting information from the Internet, the team members' work, or from their own individual ideation as demonstrated in the sequence pattern "copy and paste (C&P)" → "rephrase or revise previous work (RE)". One significant sequence of the low-performing group was "OFFTASK" → "discuss the topic (DIS)", which infers that the students tended to discuss irrelevant things or play with the system before discussing the topic.

**Figure 1.** Argumentative Knowledge Co-construction Behavior Patterns of High- (left) and Low-performing (right) Groups.

4.2 The High- and Low-performing Group’s Arguments Development
To answer the second research question about the development of the high- and low-performing groups’ arguments compositions, the Sankey diagram was employed to unpack the knowledge co-construction process to show the reasoning flow to reveal the differences between the two groups. Two Sankey diagrams were derived to visualize the development of students’ argument construction (see Figure 2) from the individual ideation phase to the group synergy phase. The left column of the diagrams is the students’ individual ideation whereas the right column is the results of group argumentative knowledge co-construction. The middle column illustrates how each idea and argument element developed. For example, the evidence (E) in Argument 4 (left column) became the evidence (E1) in Argument 4 (right column) through synthesizing the original evidence (E) with new ideas from the Internet, the team members’ work, or from their own individual ideation.

Figure 2 shows that the high-performing group students constructed eight arguments from seven different perspectives, including only two sound and completed arguments with a "C-E-R" structure (Argument 4 and 5), three arguments consisting of a claim and one piece of evidence (Argument 6, 7 and 8), and three arguments with one claim (Argument 1 and 2). After the knowledge co-construction, there were three arguments with a "C-E-R" structure and two of the three arguments consisted of richer evidence and stronger reasoning (Argument 4 and 7). Additionally, two arguments were made up of a claim and a piece of evidence (Argument 1 and 2), and in Argument 8 there were a claim and two lines of reasoning. The arguments focusing on society and technology appeared in individual ideation passe were deleted after the group knowledge-co-construction.

The low-performing group started with three completed arguments (Argument 1, 3, and 5) from perspective of society, art, and environment; two arguments with one claim (Argument 2 and 4); one argument with a claim and a line of reasoning (Argument 6). Three ideas posted were not considered as arguments by the researchers as they did not correspond to any certain claim. At the end of group co-construction, four fragmented ideas remained and there were only two sound arguments with a "C-E-R" structure (Argument 1 and 3) which was less than their individual ideation work, though Argument 3 added one line of reasoning. Through rephrasing and synergizing, some ideas were still not integrated into an argument well.

The results from the Sankey diagrams show that there were differences in the process of argument development between the high- and low-performing groups. From the behavior point of view, the high-performing group contributed new argument elements via creating reasoning and evidence (N=11) during the argumentative knowledge co-construction, but the low-performing group deleted nine elements and created only one element (R2) and a fragmented idea (idea 6). Second, from the perspective point of view, the high-performing group considered all the perspectives they proposed during the individual phase whereas the low-performing group paid most of their attention to the perspective of art only.
4.3 The Social Interaction Network of the High- and Low-performing Group

Social network analysis (SNA) was used to analyze the social characteristics of group interactions through online collaboration. This study focused on the frequency of interaction, the frequency of built-on work a participant makes to others. The degree centrality and Group centralization to show whether the network is equal or not (Yustiawan et al., 2015); In addition, the density and connectedness of each group were analysed to show the cohesiveness and reachability between group members.

The results show that the high-performance group had higher interaction frequency (freq. = 5, M = 2.5, SD = 1.732) than the low-performance group (freq. = 4, M = 2, SD = 0.816). The high-performance group also has a higher average degree (outdegree = 1.25, indegree = 1.25) than the low-performance group (outdegree = 1, indegree = 1). The high-performance group has lower out-centralization (out-centralization = 0.111, in-centralization = 1) than the low-performing group (out-centralization = 0.444, in-centralization = 0.889). On the contrary, the low-performance group has higher density and connectedness (density = 0.333, connectedness = 0.333) than the high-performing group (density = 0.25, connectedness = 0.25).

5. Conclusions, Discussion, and Limitations

This study explored the high- and low-performing groups’ argumentative knowledge co-construction from epistemic, argument, and social perspectives. The findings from LSA, Sankey diagram visualization, and SNA show that the high-performing group possesses a higher level of engagement, cognitive elaboration, and equal relationship, and is more competent in integrating argumentation elements than the low-performing group.

This study yields two findings on the epistemic dimension of collaborative argumentation. First, the high-performing group was more engaged in the learning task than the low-performing group, which echoed the findings from the literature that the more engaged group members are in the ongoing and constant elaboration of collaborative tasks, the higher the quality of learning they achieve (Barron, 2003). Second, the high-performing group tended to process information through rephrasing and synergizing pieces of information from the internet and existing ideas to form higher levels of reasoning.

On the contrary, the low-performing group tended to copy information directly from the internet without deep information processing. A possible reason for this finding is that the less successful groups tend to possess a lower level of cognitive elaboration competence. Cognitive elaboration, a crucial type of cognitive processing with regard to knowledge
construction (McNamara et al., 1996), is the enrichment of learning material using additional
information taken from or inferred in combination with prior knowledge. Arguments with high
quality in terms of argumentative knowledge come with in-depth cognitive elaboration
(Stegmann et al., 2007). This finding is consistent with previous research studies which
suggested that deep cognitive elaboration of the learning material is causally related to
knowledge acquisition (e.g., Stein & Bransford, 1979). Therefore, future pedagogical design
that aims to improve students’ argumentation skills can pay more attention to students’
cognitive processing and elaboration.

There are two findings derived from the analysis of the argument dimension. The
argument structure of the high-performing group became stronger and richer, but it was hard
for the low-performing group to integrate fragmented reasoning and evidence to support a
corresponding claim so that a complete argument could be formed. The findings are consistent
with Sampson and Clark’s (2011) research results: high-performing triads constructed
arguments that included a sufficient and accurate explanation that was well supported by
appropriate evidence and reasoning, whereas the low-performing triads produced arguments
supported by inappropriate and inadequate justification. A possible reason is that the low-
performing group students have difficulties in linking evidence to claims (Osborne et al., 2004;
Berland & Reiser, 2009; McNeill, 2011). As Osborne et al. (2004) observed, constructing a
good argument is not a simple task for school students, and they need sufficient support and
clear guidance to build their sense of what an effective argument is. Walker et al. (2019) also
reported that students may need additional guidance when they intend to revise or change
their claims and when they use scientific principles in their justification. Researchers and
teachers would need to provide additional scaffoldings to guide students to construct effective
arguments.

The results of the social interaction show that high-performing group members
contributed actively and equally to their group work, while the contribution of low-performing
group students was less equal with two students dominating the group work. Borge et al. (2012)
reported a similar case that the greatest difference between high and low-performing groups
was how dominance shifted by person across activities. For example, in high-performing
teams, while one person dominated synthesis-related activities, another person dominated
activities involving information interpretation or decision-making. In contrast, a low-performing
team would have the same person dominating both activities. This finding can be supported
by the previous research studies which showed that an equal and distributed collaborative
relationship with minimal hierarchical control could encourage students to engage more in the
collaborative process (Zhang et al., 2009). It is necessary to encourage students to build
collective responsibility (e.g., shared roles of knowledge builders) and both self- and shared-
regulation skills (Park et al., 2019).

This study has some limitations. This study is exploratory in nature with only one high-
and low-performing group analyzed. The time duration of the study was short due to the
school’s curriculum tie constraint. Therefore, the findings may not be generalizable to other
groups and contexts. Nevertheless, the study is innovative in its analytical methods and
provides interesting insights into the different knowledge co-construction processes between
the high- and low-performing groups. Future research studies will be conducted with more
groups of data included and longer duration and deeper engagement of students’ knowledge
co-construction are designed and implemented.

Acknowledgements

This study was funded by Singapore Ministry of Education (MOE) under the Education
Research Funding Programme (OER 17/19 CWL) and administered by National Institute of
Education (NIE), Nanyang Technological University (NTU), Singapore. The study was
approved by the NTU IRB (IRB-2020-04-031). Any opinions, findings, and conclusions or
recommendations expressed in this material are those of the author(s) and do not necessarily
reflect the views of the Singapore MOE and NIE.
References


Discussion Support Agent System to Promote Equalization of Speech among Participants

Ryunosuke NISHIMURA\textsuperscript{a*}, Risa IHARADA\textsuperscript{a}, Yuya SUGAMOTO\textsuperscript{b}, Yutaka ISHII\textsuperscript{b}, Toshio MOCHIZUKI\textsuperscript{c} & Hironori EG\textsuperscript{a}

\textsuperscript{a}Department of Informatics, The University of Electro-Communications, Japan
\textsuperscript{b}Department of Systems Engineering, Okayama Prefectural University, Japan
\textsuperscript{c}School of Network and Information, Senshu University, Japan

\textsuperscript{*} r.nishimura@uec.ac.jp

Abstract: Reducing the imbalance of speech length in discussions is expected to improve decision making and collaborative learning. In this paper, we propose a facilitation system based on the cumulative percentage of utterances in discussions. We surveyed the effectiveness of this system on the collaborative learning from the perspective of standard deviation of utterance length and percentage of utterances throughout discussions. We collected 20-minutes discussion data under the topic "What is a thoughtful consideration?" and "What is happiness?" from seven groups of three participants. Each group had two discussions, one with the system and one without the system. The results revealed that the facilitation system decreased subjective and objective imbalance of speech length. Additionally, turn-taking was active in the group which facilitation by the system was effective.

Keywords: Discussion support agent system, dialog analysis, facilitation, collaborative learning, percentage of utterances

1. Introduction

Recently, small group discussions are used in collaborative learning to avoid social loafing (Latané et al., 1979). Especially, researchers have focused on supporting collaborative learning with technology (DeSanctis & Gallupe, 1987; Alavi, 1994; Jermann et al., 2001). In collaborative learning, persons are required active participation because persons discuss different ideas and consider these ideas carefully.

However, the imbalance of speech length may occur because of status of the discussion such as the relationship and communication skills of participants. In addition, decision making in discussions are affected by talkativeness (DiMicco et al., 2007). This suggests that the approach to support collaborative learning and increase the productivity of the discussion by equalizing utterance length. Ishikawa et al (2019) shows that participants in discussion can learn co-regulation strategy using agent facilitation based on the percentage of utterances up to 30 seconds before facilitation. On the other hand, this system doesn’t have enough effects to equalize the percentage of utterances. If a person speaks much in 30 seconds before the facilitation, that person speak much throughout the discussion.

In this study, we proposed a system for facilitation using the cumulative percentage of utterances in discussions. To consider percentage of utterances throughout discussions, our system can facilitate discussions with less influence from the imbalance of utterances in a short time interval. From the above, we focus on two research questions (RQ):

RQ1: Can the facilitation method using the cumulative percentage of utterances throughout the discussion equalize the percentage of utterances?
RQ2: If the method is effective (or not effective), how the system reduced (or increased) utterance imbalance?
2. Related works

Commonly methods to reduce the number of utterances is that to visualize the imbalance of participation. Bachour et al. (2010) proposed the system for avoiding over and underparticipation to indicate how much the individual has spoken. In this study, Matrix of LEDs indicate person’s participation levels based on the number of speaks. Similarly, Kim et al. (2008) visualized the participation as a figure consisting of a circle representing the imbalance of the utterances and some lines representing the amount of each participant’s speech. In addition, Bergstrom & Karahalios (2007) examined the expression of the Conversation Clock which visualize prosodic features such as a period of silence and volume of speech. There are also ways to express discussion structure using metaphors such as trees and plants (Tausch et al., 2014; Praharaj et al., 2019).

In these methods, all participants can check the discussion status such as the speech imbalance displayed on the table. However, this may be uncomfortable for the reluctant speaker because everyone knows that he or she is not contributing to the discussion. Therefore, we believe that it is suitable to encourage the participation of the speakers by facilitating them while the agents maintain the audio features. One such approach is using robots (Miyake et al., 2012). However, it may be difficult to prepare robots for actual discussions because they are expensive. In anticipation of such application to actual discussions, the facilitation agent system proposed by Ishikawa et al. (2019) was used. The facilitation system can be used in a variety of settings as long as there are displays and polarizers to project the agents, microphones, and computers to run the system.

3. A Facilitation system based on the percentage of utterances throughout discussions

3.1 Overview of the facilitation system

Figure 1 shows the system configuration. Our system consists of two parts: discussion analysis part and agent control part. The discussion analysis part calculates the speech imbalance based on the speech collected by the wearable speech collection device and sends a request to the agent control part according to the degree of the imbalance. The agent control part operates the holographic agent using request sent by the discussion analysis part.
3.2 Calculating the cumulative percentage of utterances

The cumulative percentage of utterances is the rate of each participant's speech when the total time of all participants' speech from the start of discussion to the time of the calculation is set to 100%. Silent periods are not included in the calculation of the percentage. For instance, if three participants speak for 2, 3, and 5 minutes respectively in a 10-minute discussion, the cumulative percentage of utterances is 20%, 30%, and 50%, respectively.

3.3 Criteria for agent facilitation

The holographic agent facilitates discussions when the system detects that a speech imbalance is occurring in a discussion. In general, when the number of participants in a discussion is \( n \), if the cumulative percentage of utterances in all participants is \( 100/n\% \), it can be considered that there is no imbalance in the speech of each participant. However, it is unlikely that the cumulative speech rate of all the discussion participants coincides. Therefore, we set \( 100/n\% \pm (100/n - 100/(n+1))\% \) as the acceptable range, and the situation in which there is a participant who is out of this acceptable range as the biased state. For example, for a three-person discussion, the acceptable range would be 25% to 41%.

The agent suppresses the speech of the participants whose cumulative percentage of utterances is higher than the acceptable range, and encourages the speech of the participants whose cumulative percentage of utterances is lower than the acceptable range. When the cumulative percentage of utterances of all participants is within the acceptable range, the agent does not approach them, but listens or observes them. Therefore, the following three facilitations are performed by the agent system. Decisions on agent behavior based on the cumulative percentage of utterances were made every minute.

- If there is one or more discussion participants whose cumulative percentage of utterances is lower than the acceptable range and there is no discussion participant whose cumulative percentage of utterances is higher than the acceptable range.

In this case, there is an imbalance in speech because there is a participant who does not speak much. Therefore, the agent asks the participant to "Why don't you express some opinions?" or "Why don't you explore the opinions so far?" with voice.

- If there is one or more discussion participants whose cumulative percentage of utterances is higher than the acceptable range and there is no discussion participant whose cumulative percentage of utterances is lower than the acceptable range.

In this case, there is a imbalance in speech because there is a participant who speaks much. Therefore, the agent asks the participant to "Why don't you ask others?" or "Why don't you explore other people's opinions?" with voice.

- If there is one or more discussion participants whose cumulative percentage of utterances is higher than the acceptable range and lower than the acceptable range.

In this case, there is a imbalance in speech because there are both a participant who speaks much and speaks little. In this case, the above two cases are alternated.

4. Experiment

An experiment was conducted to evaluate the effectiveness of the facilitation system. Each participant had two discussions, one is the experimental discussion using the system, and another is the control discussion without the system. The order of the two discussions was
Two discussion topics were set as "What is thoughtful consideration?" and “What is happiness?". Discussion topics are based on a course of study in Japan (MEXT, 2020). These discussions were attended by three graduate and undergraduate students of a science and engineering university. Each discussion had three participants and lasted for 20 minutes.

After discussion, we conducted a questionnaire to evaluate the system. In the system evaluation questionnaire, participants were asked to describe how they felt the system facilitation. Figure 2 shows the discussion with the system.

![Figure 2. A discussion with the agent](image)

5. Results

5.1 RQ1: Can the facilitation method using the cumulative percentage of utterances throughout the discussion equalize the percentage of utterances?

We examined the effects of the facilitation system on equalization of speech from the viewpoint of the standard deviation of the cumulative percentage of utterances. This standard deviation represents the degree of speech imbalance among the participants. That is, if the standard deviation in the experimental discussion is smaller than the standard deviation in the control discussion, the system should promote equalization of speech among participants. Table 1 shows the standard deviation of the cumulative percentage of utterances in seven groups.

<table>
<thead>
<tr>
<th>Group No.</th>
<th>Control discussion</th>
<th>Experimental discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.44</td>
<td>2.23</td>
</tr>
<tr>
<td>2</td>
<td>13.22</td>
<td>4.63</td>
</tr>
<tr>
<td>3</td>
<td>13.52</td>
<td>6.36</td>
</tr>
<tr>
<td>4</td>
<td>17.58</td>
<td>13.40</td>
</tr>
<tr>
<td>5</td>
<td>25.45</td>
<td>15.43</td>
</tr>
<tr>
<td>6</td>
<td>23.85</td>
<td>18.06</td>
</tr>
<tr>
<td>7</td>
<td>20.66</td>
<td>19.56</td>
</tr>
</tbody>
</table>

5.2 RQ2: If the method is effective (or not effective), how the system reduced (or increased) utterance imbalance?

From Table 1, the method should be effective to promote equalization speech among participants. However, the degree of mitigation of the speech imbalance by the facilitation system was different. Therefore, we divided the seven groups into two groups: a large effect...
group (group 1-3) and a small effect group (group 4-7). Additionally, we examined the
process of equalization in the large and small effect groups. Figure 3 shows the percentage
of utterances per minute in the experimental discussion of Group 3 and Figure 4 shows that
of Group 7. Group 3 had the highest number of the facilitation in the large effect group and
Group 7 had the largest the standard deviation in the small effect group.

The percentage of utterances is recalculated every minute in Figure 3 and Figure 4. For
example, Figure 3 shows that participant 1 spoke longer than other participants in the
interval from 60 to 120 seconds.

6. Discussion

Table 1 shows standard deviation of the cumulative percentage of utterances decreased for
all groups to a certain degree. Therefore, RQ1 is partially shown. In addition, Figure 3 and 4
suggest that turn-taking differs between the large and small effect groups. For example,
each cumulative percentage of utterances per minute in Group 3 is equalized and cross
many times after the instructions in both the first and second halves of the discussion. This
means speaking turn rotated fast and equally in Group 3. On the other hand, each
cumulative percentage of utterances per minute in Group 7 is not equalized and cross few
times despite being facilitated by the agent system. Similar results were observed in the
questionnaire. Participants in Group 3 answered that they were aware of speech imbalance
due to the facilitation, while participants in Group 7 answered that they ignored it due to the
difficulty of following the facilitation. Therefore, awareness of turn-taking is the key to
answering RQ2: in order to equalize speech, it is necessary for participants to be aware of
speech imbalance through facilitation and to correct it actively by turn-taking. These results
imply the effectiveness of facilitation may be evaluated by observing the turn-taking after the
facilitation.

7. Conclusion

In this paper, we proposed the facilitation system based on the cumulative percentage of
utterances in discussions. Here, the experiment yields two important results: this system is
contributed to promote equalization of speech among participants and turn-taking is a key to
equalize cumulative percentage of utterances. Our study gives a quantitative perspective of
the facilitation.
However, this study has some limitations. In this study, we mainly analyzed large
effect group and small effect group from the graph of cumulative percentage of utterances.
We need to further explore factors that may make the difference in the effectiveness of the
facilitation system such as personality characteristic.

Acknowledgements

This work has been partly supported by the Grants-in-Aid for Scientific Research (NOS.
21K02752, 21K18527 and 22K02951) by MEXT (Ministry of Education, Culture, Sports,
Science and Technology) in Japan

References


Bergstrom, T., & Karahalios, K. (2007, January). Conversation Clock: Visualizing audio patterns in co-
located groups. In *2007 40th Annual Hawaii International Conference on System Sciences
(HICSS’07)* (pp. 78-78). IEEE.


face-to-face collaborative discussion. In *Collaboration Technologies and Social Computing: 25th
International Conference, CRIWG+ CollabTech 2019, Kyoto, Japan, September 4–6, 2019,
Proceedings 25* (pp. 212-221). Springer International Publishing.

art technology for supporting collaborative learning. In *European Perspectives on Computer-
Supported Collaborative Learning* (pp. 324-331).

group collaboration using sociometric feedback. In *Proceedings of the 2008 ACM conference on
Computer supported cooperative work* (pp. 457-466).

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and
consequences of social loafing. *Journal of personality and social psychology*, 37(6), 822.

MEXT (Ministry of Education, Culture, Sports, Science and Technology). (2020, October). English
Translation of the Revised a Course of Study at Junior High Schools Revised in 29th year of the
Heisei period (provisional translation). Retrieved from https://www.mext.go.jp/content/20201008-
mxt_kyoiku02-000005242_1.pdf.

Miyake, N., Okita, S., & Rose, C. (2012). Robot facilitation as dynamic support for collaborative
learning.

Collaboration. In *Transforming Learning with Meaningful Technologies: 14th European
Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands,

supporting brainstorming through a metaphorical group mirror on table or wall. In *Proceedings of
the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (pp. 541-
550).
Experimental Verification of "Peer-ness" Formation by a Learning Companion Robot
—Possibility of inducing a sense of competition through long-term nonverbal interaction—

Koki Honda\textsuperscript{a}, Yoshimasa Tawatsuki\textsuperscript{b} & Tatsunori Matsui\textsuperscript{c}
\textsuperscript{a}Graduate School of Human Sciences, Waseda University, Japan
\textsuperscript{b}Center for Data Science, Waseda University, Japan
\textsuperscript{c}Faculty of Human Sciences, Waseda University, Japan
*kouki0506hd@akane.waseda.jp

Abstract: The purpose of this study is to construct a companion robot behavior model that provides a learning environment enhancing learners' motivation to learn by inducing peer awareness. Particularly, inducing "peer-ness" through robot behavior has an impact on education. Therefore, we verified a robot behavior model oriented to induce "peer-ness," including inducing a "sense of competition," through long-term nonverbal interactions. The results of this study confirmed the "Peer-ness" induction and demonstrated the validity, although it varied depending on the mental state.

Keywords: Peer learning, learning companion robot, non-verbal-communication

1. Introduction
With the development of information technology, the GIGA school concept was launched in Japan. Consequently, a form of teaching and learning known as e-learning has become more common. In addition, a rapid increase in the demand for e-learning was observed during the COVID-19 pandemic. However, the form of learning support in e-learning is dominated by support from a learner who is more knowledgeable about the learning domain than the learner who is better suited to transferring knowledge and facilitating understanding. However, there exists significant learning (peer learning), in which knowledge is consolidated and confirmed through collaboration with classmates. This learning, represented by active learning, is flourishing in school environments. Therefore, although Peer Learning is not conducted in on-site e-learning environments, it is limited to creating a single-person learning environment. However, Peer Learning has a positive impact not only on learning effectiveness but also on learner psychology (Marder, et al. (2017)). Therefore, we aim to construct an e-learning environment that brings about a positive psychological effect of Peer Learning by using a learning support robot and investigate the possibility of inducing a sense of camaraderie (peer-ness) through the robot.

1.1 Peer Learning and Learning Support Robots
Research has been conducted on learning support robots. Kashihara (2019) showed that, unlike other media, learning support robots, which have physicality, can achieve more natural and authentic communication with learners. Hence, it is expected to reduce negative emotions such as learning inhibition. In a study using a computer as a companion in a learning environment, Kasai et al. (1999) developed a multi-agent learning environment in a cooperative learning environment that supports learning by using the companion as either a peer or teacher, depending on the learner's situation. A multi-agent learning environment was developed in a cooperative learning environment. Chan (1991) pointed out that a
companion robot that behaves as a teacher may give the impression of being "watched." However, Marder et al. (2017) suggested that the presence of a peer can produce positive effects. In general, while previous studies on peer learning and learning support by robots have suggested their usefulness, they have not considered the feeling of "being monitored" that a teacher-like position brings. Therefore, this study aims to construct a "robot behavior model that induces peer-ness," oriented toward the construction of robots that exhibit only appropriate peer positions in response to various situations.

1.2 Intrinsic Motivation and Nonverbal Information

Numerous studies have examined the positive effects of intrinsic motivation on learning. Regarding autonomy-induced behavioral decisions related to intrinsic motivation, Yokoyama et al. (2009) stated that others' intention estimations influence one's own behavioral decisions. Osawa (2014) found that the behavior of the target induces intention estimation; anthropomorphism is used to superimpose the limb movements of the robot behavior on the movements of human limbs, and intention estimation from the target's behavior is performed by anthropomorphizing the target. Our study aims to boost learner autonomy through intentional inference from a robot's behavior and foster intrinsic motivation using nonverbal anthropomorphic cues. The robot's behavior model excludes linguistic information and relies solely on nonverbal cues, including nonsense utterances.

1.3 Short-term study of induction of "Peer-ness" by robot motion

We have experimentally clarified the mental factors that constitute the Peer-ness of learners during learning and created a psychological model with a hierarchical structure (Figure 1)(Koki (2022)). The meaning of each factor is listed in Table 1. The higher the evaluation value of each, the higher the impression of the psychological factors that exist at the top, and the more "Peer-ness" is felt in the end.

![Figure 1 Psychological model of feeling peer](image)

<table>
<thead>
<tr>
<th>factor</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense of Same</td>
<td>Feel as if companion is in the same situation and condition as you</td>
</tr>
<tr>
<td>Sense of Competition</td>
<td>Feel as if companion is competing with you</td>
</tr>
<tr>
<td>Sense of Trust</td>
<td>Feel that the companion is trustworthy</td>
</tr>
<tr>
<td>Sense of Affinity</td>
<td>Feel the closeness</td>
</tr>
<tr>
<td>Sense of Equivalence</td>
<td>Feel you and companion are of equal ability</td>
</tr>
<tr>
<td>Sense of Urgency</td>
<td>Feel an imminent gap in ability between you and companion</td>
</tr>
<tr>
<td>Insecurity of Position</td>
<td>Feel the state of being pulled out - pulled in</td>
</tr>
<tr>
<td>Sense of Awareness</td>
<td>Feel somewhat concerned about it</td>
</tr>
<tr>
<td>Sense of Comprehend</td>
<td>Feel you and companion understand each other.</td>
</tr>
<tr>
<td>Sense of Respect</td>
<td>Feel that companion has superior capabilities</td>
</tr>
</tbody>
</table>

In an attempt to induce "Peer-ness" by robot actions according to the created psychological model, we investigated the relationship between robot actions and the psychological model, and based on the obtained results, we conducted a short-term study on inducing "Peer-ness." However, no strong results were obtained for inducing "Peer-ness" based on the
"sense of competition" and "sense of competition," which are severely generated instantaneously, suggesting the need for a long-term study.

2. Method

This long-term study aims to induce a sense of competition and peer-ness, building on the findings of short-term studies. Owing to data limitations, the participants included one male junior high school student with basic learning abilities and four university students (two men and two women) commuting to Tokyo. This study focuses on the results of an experiment with junior high school students. The robot behavior and psychological model used were based on previous research (Koki, 2022), which led us to conduct a long-term interaction impression evaluation experiment with junior high school students with diverse developmental abilities in Tokyo. Referring to Kanda et al. (2004), we created episodes showing robot behavior corresponding to situations that were assumed to induce a "sense of competition," based on the data obtained in previous experiments. Additionally, the robot was instructed to perform mathematical learning in conjunction with it. The experiment was conducted three times a week for two weeks for one hour per day, referring to Marder et al. (2017) and Kanda et al. (2003), who studied long-term interactions. Figure 2 illustrates the experimental flow and Table 2 presents the details.

![Figure 2: Experimental flow](image)

### Table 2. The details of the experiment

<table>
<thead>
<tr>
<th>Session</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>Introduce about this experiment</td>
</tr>
<tr>
<td>First Research</td>
<td>Research about character (age, sex, motivation of math study, impression of robots [7]) and skill of math</td>
</tr>
<tr>
<td>Study</td>
<td>Three sessions in one study</td>
</tr>
<tr>
<td></td>
<td>In one session, the robot performs (cause sense of competition) after one mini test</td>
</tr>
<tr>
<td></td>
<td>The test has 3 levels (low-middle-high) decided by the result of First Research</td>
</tr>
<tr>
<td></td>
<td>First session: low level test</td>
</tr>
<tr>
<td></td>
<td>Second session: high level test</td>
</tr>
<tr>
<td></td>
<td>Last session: middle level test</td>
</tr>
<tr>
<td>On-time Questionnaire</td>
<td>Administered immediately after the experiment to assess impressions of &quot;Sense of Competition&quot; and &quot;Peer-ness&quot;, willingness to learn, and impressions</td>
</tr>
<tr>
<td></td>
<td>“Sense of Competition” and “Peer-ness”: Evaluation by Vas method and Comparison with the previous study</td>
</tr>
<tr>
<td>Final Research</td>
<td>A questionnaire survey</td>
</tr>
<tr>
<td></td>
<td>Administered to learners, asking them to rate their impressions of &quot;competitiveness&quot; and &quot;peer&quot; as well as their feelings about learning on a scale of 1 to 10</td>
</tr>
</tbody>
</table>

In addition to the experimental flow, interviews were conducted at the end of each day to gather the participants' impressions of the robot and their learning. The study focused on individual learning in an e-learning environment with the subjects and experimenter in

274
separate spaces divided by an impulse screen. To maintain a sense of solitude during the study, learners received instructions from a third person and worked alone, answering exercises on loose-leaf paper while being monitored using a wearable camera. The actions of the robot were presented using the Woz method based on the episodes created.

![Image of the experiment setup]

Figure 3: Schematic diagram of the experiment and environment

3. Result

3.1 Relationship between “Peer-ness” and “Sense of Competition”

The graphs illustrate the transition of the impressions of "peer-ness" and "competitiveness," respectively. The vertical axis indicates the score obtained using the VAS method converted to a perfect score of 100, and the horizontal axis indicates the number of experiments performed. The experimental days are indicated by red lines in figure 4.

![Graphs showing impressions over time]

Figure 4: Impression evaluation results for "Peer-ness" (L) and "Sense of competition" (R)

Figure 4 indicates that the “Peer-ness” impression score remained high throughout all sessions, and the effect on “sense of competition” was unclear. However, “Peer-ness” score was notably lower at the 8th and 9th times, coinciding with participants expressing confusion about robot actions during interviews. After the experiment, participants mentioned feeling more similar to their peers, suggesting the potential induction of “peer-ness” during the study. Graphs indicate increasing impressions of “peer-ness” and “competitiveness” with each experimental day. This suggests a positive correlation between “competitiveness” and “Peer-ness” increase over time, though the connection throughout the day was not consistently strong.

3.2 Motivation to Learn Mathematics

This section describes the changes in the participants' willingness to learn mathematics observed throughout the experiment. In the pre-survey, participants were asked to rate their willingness to learn mathematics on a 10-point scale. The initial score was eight (on a 10-point scale). However, in the last day's interview, the students reported that they tended to give up without solving written answers in the early days. Hence, as an initial situation, willingness to learn mathematics depends on the problem being addressed. The following describes the content of the questionnaire feedback section from the last session of the experiment.

I really want to do math problems now!
It was fun to solve the problems so easily.
I was happy that the robot was kind to me.

![Feedback from the last session]

Figure 5: Impressions after the 10th session
This suggests that the students were more motivated to learn mathematics throughout the experiment. In addition, to confirm their reflections throughout the experiment, we focused on the contents of the feedback column in the final survey. The actual content is presented below.

<table>
<thead>
<tr>
<th>I used to think that robots were something that would one day destroy humans (due to anime and Kamen Rider). But now I feel that robots can get along well with humans. I have grown attached to robots. Solving computational problems has become more fun than before.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 6: Contents of the Final Research Comments Section</td>
</tr>
</tbody>
</table>

These results suggest the aforementioned suggestion of increased motivation is stronger.

### 3.3 Change in Robot Impressions

In this experiment, the participants were asked to answer questions about their impressions of the robot in a questionnaire survey after each experiment and in an interview survey after each experimental day. The results revealed that after repeated interactions, the participants began to judge the robot as a human being. Below are the results of the first and fourth interviews, and the results of the interview on the third day (after the fifth interview), in which changes were particularly evident.

<table>
<thead>
<tr>
<th>The applause for the first robot made me very happy. Maybe that's why I was a little disappointed that the second time it did so-so. However, I was very happy when the robot gave me a &quot;○&quot; at the end. I'd like to solve the 2nd question one more time because I think I can solve it. I think I can do it.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 7: Impressions after the first session</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The first round of applause seemed to be done by an adult. It was as if they were saying, &quot;Mm, well done.&quot; The second round of applause looked like a friend on my side. It seemed as if he was saying, &quot;Well done.&quot; The third time, he was embarrassed (?). The third time, she was making an embarrassed (?) motion, but I couldn't understand it well.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 8: Impressions after the fourth session</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(Test subject) I guess so...the robot is starting to look more and more like a human after all. (Experimenter) Human...I see... (Test subject) I see... I was surprised that the robot was moving at first, but now it seems normal and natural...yes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 9: Day 3 Interviews</td>
</tr>
</tbody>
</table>

The results of the interviews indicate that, for the first time, the learners began to see the robot as a human being, something that had not occurred before the second day of training. The results of the questionnaire survey indicated that the robot began to guess what it wanted to say to learners.

### 4. Discussion

The following two points are important inferences from the results of this experiment.

- There is a proportional relationship between the "sense of competition" and "Peer-ness" ratings over the course of a day.
- Induction of intention estimation by presentation of nonverbal information
4.1 A Proportional Relationship between the "sense of competition" and "Peer-ness" Ratings over the Course of a Day

The results of this experiment indicate that there was no improvement in each of the "competitiveness" and "peer-ness" ratings throughout the entire 6-day experiment. However, the results of the experiment indicated that the two factors improved proportionally. This suggests that the robot's nonverbal information may induce a sense of competition as well as peer-ness, however, it also suggests the possibility of a decrease in the evaluation of the robot as days passed. This could be due to the effects of day-switching, the cause of which remains unclear. Therefore, the effects of switching days should be further investigated.

4.2 Induction of Intention Estimation by Presentation of Nonverbal Information

The questionnaire and interview survey results indicated an increase in participants' motivation to learn mathematics throughout the experiment. This increase may not be solely attributable to improved self-efficacy through repeated learning, as problems of varying difficulty were consistently presented with some remaining unanswered within the time limit in each session. Additionally, Figures 7 and 8 depict a growing perception of the robot as human, suggesting a thought process akin to "anthropomorphism" based on Osawa (2014). The results indicated that the robot infers its intentions through the exhibited movements. Moreover, the robot's behavior starts from the learner's exercise-related information, potentially triggering an active strategy and decision-making related to the "autonomy" component of intrinsic motivation. This increase in autonomy may lead to enhanced intrinsic motivation. Additionally, learners' self-affirmation increased when they believed they could solve problems the next time, further supporting their intrinsic motivation.

References


Study on The Development of Computational Thinking Decomposition Strategies for Senior Primary Students

Mengtao LI, Yaxin GUAN & Guang CHEN
Faculty of Education, Beijing Normal University, China
*guang@bnu.edu.cn

Abstract: As an important component of Computational Thinking, the decomposition strategy has attracted the attention of researchers, but few studies have explored how the decomposition strategy itself is formed and developed. In this study, we carry out thematic-based programming teaching for senior primary school students. The research is divided into four stages: pre-test, exercise, post-test, and migration. We use Microgenetic Design to repeatedly and deeply observe students' programming learning activities in four stages, and conduct interviews and records, so as to analyze the development path and migration of students' decomposition strategies. The research results show that the decomposition ability of senior primary school students can be improved in programming learning, and the overall performance is that advanced strategies gradually replace low-level strategies. However, this process varies from person to person and is not a linear development, and there may be other influencing factors that lead to the return of low-level strategies.

Keywords: Computational Thinking, decomposition strategies, programming, primary students

1. Introduction

In the 21st century, Computational Thinking (CT) has become a prerequisite skill for individuals to cope with a wide range of new challenges and is influencing various fields related to the natural and social sciences. CT has gained worldwide attention as a key literacy for sustaining national competitiveness and has been adopted as a generic skill to be learned by all students in many national initiatives, curricula, standards and other policy documents. Since the introduction of CT, there has been a lot of focus on how to develop this skill (Hsu et al., 2018). In terms of developmental approaches, Hromkovič (2006) and Wing (2006) suggest programming as the main strategy for developing CT in schools. Scherer et al. (2019) also found through meta-analysis that programming is a primary method for developing students' CT. Nevertheless, a large amount of research has focused on the strategies and effects of developing CT, and there is a lack of research on students' strategy development and behaviour during programming.

Based on the above background, this study explored the developmental patterns of cognitive strategies during programming in upper primary school students. Focusing on decomposition strategies, which are the most difficult to be mastered (Selby, 2015) and the most fundamental (Rijke et al., 2018) in CT, we used a Microgeneration Design to describe the developmental pathways and degree of transfer of students' decomposition strategies during programming activities. On the one hand, we focus on the developmental process of decomposition strategies in primary school children's CT; on the other hand, we provide support for future insights into aspects of computational thinking and its internal structure.

2. Literature Review
2.1 Computational Thinking

CT is a term first used by Papert (1980) to describe the thought process of formulating problems and their solutions in a way that can be executed by a computer. He found that learning with a programming language helped students to simulate the algorithmic logic of a computer, which in turn influenced the way they thought (Papert, 1982). CT is considered to be the psychological basis for problem-solving, designing systems, and understanding human behaviour, and can help people simplify a difficult problem into multiple problems that we know how to solve (Wing, 2006). This process not only prepares students for study in the field of computer science but also provides them with the tools and skills to approach and solve problems in different areas of knowledge (Werner et al., 2012).

2.2 Decomposition strategy

As one of the foundations of CT, decomposition strategies can deconstruct complex problems into multiple easily solvable subproblems (J. Wing, 2008). Therefore, decomposition strategies are crucial for solving complex and large real-world problems (Rijke et al., 2018).

Tsai et al. (2022) argue that decomposition and abstraction are two basic and critical factors required for the development of CT, and suggested that decomposition and abstraction are the starting point for teaching based on the developmental sequence of CT skills. Decomposition is particularly important for the development of CT skills for beginning programmers. Decomposition has always been emphasized as a part of CT as a whole, but few studies have revealed specific details about decomposition itself (Rich et al., 2019). Although researchers have defined four levels of strategies from 0-3 in terms of assessment criteria for their development, with different conceptual and descriptive definitions (Avila et al., 2019), it does not explain how students’ Disaggregation strategies develop in the process of CT development and lacks depth and guidance for practice.

2.3 Visual programming

Visualisation means expressing abstract content in an intuitive way (Rijke et al., 2018). Unlike textual programming such as Python and Java, visual programming tools provide students with the opportunity to learn computer programming concepts through visual feedback-based learning, reducing the cognitive load placed on students by complex programming grammar (Lye & Koh, 2014), and allowing students to focus on the logic and structure involved in programming rather than worrying about the mechanics of writing programs (J. M. Wing, 2006). Numerous studies have shown that it makes sense to develop decomposition strategies from the elementary school level: on the one hand, programming develops CT (Buitrago Flórez et al., 2017). On the other hand, children already have the cognitive level to understand basic computer programming concepts by the age of four (Bers et al., 2014).

During programming, students are exposed to CT, using computer science concepts such as abstraction, decomposition, and generalization to solve problems. Programming education should focus on how students can break down and categorize complex problems to solve them efficiently (Sáez-López et al., 2016). The Block-coding programming tool used in this study is the Discovery Education Coding Curriculum (DECC), which supports students in allocating limited cognitive resources to the mobilization of problem-solving strategies for students who are new to programming.

2.4 Research questions

Piaget believes that children’s cognitive strategy development pattern has four stages from low to high. In CT development, there is variability in the development of different elements in the four stages. Even if they have been trained by teachers in advanced strategies for each element, they will be in the no-strategy stage or the low-strategy stage and will not be able to
use advanced strategies comfortably due to the lack of cognitive resources. Focusing on CT can help students be able to select and apply appropriate strategies and tools to better conceptualize, analyze, and solve complex problems. However, most research has focused on issues related to learning skills, concepts and practices (Ezeamuzie & Leung, 2022; Grover & Pea, 2013), and the specific processes involved in the development of students' CT need to be further explored. This study focuses on one of the most fundamental and critical starting points in the development of CT: decomposition strategies. By implementing instructional activities that facilitate the development of CT decomposition strategies in elementary classrooms and tracking students' behaviors and characteristics in the process, this study proposes to address the following questions:

RQ1: How well do senior primary students use decomposition strategies at different stages?
RQ2: What are the developmental pathways of decomposition strategies for students in senior primary grades at different stages of the same programming activity?
RQ3: To what extent do senior primary school students transfer different strategies within the same programming activity?

3. Research Methodology

We explored the development of CT decomposition strategies and pathways in primary school students: a micro-incidence design was used to collect and code relevant data through classroom behavioural observations and post-classroom interviews.

3.1 Subjects of study

The subjects of the study were 5th-grade students in a primary school in Beijing. Through pre-experimental observations, interviews and data analyses, 18 subjects were finally identified. During the implementation of the activity, 4 students did not participate in the complete experimental process for personal reasons, so 14 valid data were received after the teaching activity. The sample of 14 consisted of 4 boys and 10 girls, all aged 9-11 years (M = 10.10, SD = 0.59).

3.2 Study design

Microgenetic Design is a research method to study the cognitive and behavioral development process of children, which is widely used in the research of cognitive strategy development of young children to high school students (Siegler & Crowley, 1991). Through iterative and in-depth observation and analysis of changes in behavior throughout the process, Microgenetic Design can provide detailed information about the sources, pathways, velocity, and extent of cognitive changes, as well as the diversity of change patterns. There are several reasons: (1) observations cover the entire interval from the onset of change to relative stability; (2) the frequency of observations is highly consistent with the rate of change of the phenomenon; and (3) the observed behavior is finely iteratively analyzed in iterative experiments to facilitate speculation on the processes that produce qualitative and quantitative changes. In short, the high density of fine-grained data provides a guarantee of the validity of this approach.

Traditional pre- and post-test experiments have a low density of observation of the process of behavioural change throughout the study and do not allow for a sophisticated analysis of strategy development behaviour during this process. This can be well avoided by using a Microgenetic Design to observe students' behaviour and thinking in the classroom during the course of the lesson. When episodic behaviours were more ambiguous or absent, a categorical basis for strategy assessment was obtained by promptly asking students how they designed the procedure.
3.3 Research process

The experiment consisted of two parts, a pre-experiment and a formal experiment. All six sessions were taught by the same researcher and lasted two and a half weeks, with each session lasting 40 minutes. The six class periods were centred around six thematic activities that were taught as CT activities: Movement, Simple Input, Different Kinds of Input, Buttons & Commands, Sequence & Animation and Conditional Events. Each theme has three different programming activities under it, and students need to complete these three programming activities and perform a programming challenge in each lesson, which together constitute the pre-test, practice, post-test, and migration phases of the experiment.

In the pre-experiment, Topic Activity 1 was used to screen students with zero foundation. After completing each task, students were tested on their decomposition skills. All 5 thematic activities in the formal experiment consist of 3 learning tasks, which are used to assess students’ CT strategy development in the pre-test, practice, and post-test phases. In addition to this, students were required to undertake the completion of a programming project as a migration test.

3.4 Data collection

Data were collected through classroom observations and interviews. Classroom observations were conducted to track and record student behaviour and development in the classroom. Interviews were conducted in the form of questions asked by the teacher and shared by the students in the classroom and were used for further analysis and refinement. Interview outlines were prepared by the researcher based on the literature and what was focused on in this study, and the interviews included students' thoughts on programming as well as their feelings about learning during the four phases.

Data were coded according to the decomposition strategy level coding table. Based on a definition of decomposition strategy development levels proposed by Avila et al. (2019), we took into account specific programming situations as well as theories related to cognitive strategies and invited two researchers in the field of CT to classify students' decomposition strategies into four levels. The Decomposition Strategy Hierarchy scale is shown in the appendix.

4. Research findings

The results of the study will depict the change routes and development of decomposition strategies in senior primary school students during the completion of programming tasks in the following three ways.

4.1 Students' overall use of decomposition strategies at different stages

As shown in Figure 1, the Roma numerals from small to large represent the four stages of pre-test, practice, post-test, and migration. In the pre-test stage, students' main decomposition strategy is the level 0 strategy, and they hardly use the advanced strategy; in the practice stage, level 1 strategies are the students' main strategies, and level 3 strategies begin to appear; in the post-test stage, level 2 and level 3 strategies become the main strategies, and level 0 strategies have been completely replaced; in the migration stage, the use of level 3 strategies reaches 50%, and the use of level 2 strategies is as high as 45%.
4.2 Development of students’ decomposition strategies in programming activities Pathways

As shown in Figure 2, the line from light blue to dark blue represents the four-level decomposition strategy from low to advanced (e.g. the lightest blue line represents the 0-level decomposition strategy). The route analysis of the strategies adopted in four different stages of the same programming activity shows that the development trend of different levels of decomposition strategies in different thematic activities is basically the same, which is manifested as the gradual replacement of low-level strategies by high-level strategies.

![Figure 2](image)

**Figure 2.** The overall line of development of the decomposition strategy under different thematic activities.

4.3 Transfer of students' decomposition strategies in programming activities extent

Figure 3 shows a roadmap for individuals to adopt strategies at different stages of programming activity.  represents the level, and the number on the horizontal line between the two levels represents the number of people whose route changes between the two strategy levels. The
results show that before being taught programming, all students initially used only low-level decomposition strategies, and the majority of students used zero-level strategies. In most cases, the decomposition strategies used by students follow a continuous development from low-level to high-level strategies, but there will be a brief regression from high-level strategies to low-level strategies.

![Diagram](image)

*Figure 3. The degree of individual migration of strategies in different activities.*

### 5. Discussion and conclusions

The purpose of this study is to investigate the developmental pathways and strategy transfer of CT decomposition strategies of students in the upper primary grades, and we can summarise three main findings. Firstly, students' decomposition strategies can be developed through topic-based programming learning, and programming instruction in the upper primary grades can be an effective way to develop students' decomposition skills. This result reaffirms Bers et al.,(2014) and Rijke et al.,(2018) that it is feasible and meaningful to develop students' decomposition skills by teaching programming from the primary level. Secondly, students' decomposition strategies are not linear but a dynamic process of development, and there may be other factors that lead to the recurrence of low-level strategies. This result is similar to the findings of Siegler & Stern(1998) on the development of mathematical strategies in second-graders. In addition, students' decomposition strategy development is a complex process, and students' own experiences, teachers' interventions, and differences in instructional content may affect the developmental trends and the degree of transfer of students' decomposition strategies. In the process of guiding students to learn programming, teachers need to provide effective guidance to help students decompose problems and tasks. At the same time, through the creation of specific problem situations and practice opportunities, teachers can transfer the decomposition strategy teaching content to other disciplines and make effective adjustments to the curriculum design, so as to improve students' decomposition strategies in multiple ways.

### Preference


Appendix

Table 1. Decomposition Strategy Hierarchy

<table>
<thead>
<tr>
<th>Grade</th>
<th>Definition</th>
<th>Specific description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>During the analysis of the programming task, programming relies on subjective judgement and may not achieve the programming objectives.</td>
<td><strong>Subjective judgement:</strong> the process of identifying the task problem and thus completing the programming by virtue of an autonomous understanding of the programming task, in which errors may occur. <strong>Missing objects:</strong> Objects needed in the programming are not added to the scene. Resulting in missing code. <strong>Missing code:</strong> Missing code in the programming process due to a lack of systematic decomposition of programming tasks.</td>
</tr>
<tr>
<td>Level 1 Strategies</td>
<td>Objective decomposition of programming tasks based on subjective judgement, using combinations of existing code to complete programming tasks.</td>
<td><strong>Task decomposition:</strong> the ability to decompose the main subtasks of a programming task. <strong>Object decomposition:</strong> the ability to program action codes for each object based on the objects that appear in the programming.</td>
</tr>
<tr>
<td>Level 2 Strategies</td>
<td>Be able to completely analyse the objectives of a task in programming, analyse the relationships between subproblems and use combinations of previously learned code blocks to complete programming tasks.</td>
<td><strong>Sub-task combination:</strong> the ability to break down the overall goal of a programming task into several task blocks and combine them using existing code to complete the programming. <strong>Sequential decomposition:</strong> problem decomposition in the order of the problems that can be programmed for a programming task. <strong>Conditional decomposition:</strong> the ability to decompose programming tasks according to the different situations of the programming problem.</td>
</tr>
<tr>
<td>Level 3 Strategies</td>
<td>The process of decomposing task objectives enables a systematic breakdown of programming objectives into smaller problems in a combination of process and content, which are effectively combined to complete the programming task.</td>
<td><strong>Autonomous decomposition:</strong> the ability to autonomously identify the main problems in a programming task of their own design and the relationships between the various sub-problems. <strong>Optimal sub-problem combination:</strong> The ability to form an optimal combination of sub-problems in programming into an efficient solution to the entire programming task.</td>
</tr>
</tbody>
</table>


Remembering the knowledge of experts and novices in computer-supported collaborative learning: A multinomial processing tree approach

Oktay ÜLKERa*, Daniel BODEMERa
aUniversity of Duisburg Essen, Germany
*oktay.uelker@uni-due.de

Abstract: Accurately assessing learning partners' knowledge profiles improves collaborative learning. Group awareness tools facilitate constructing such social context knowledge during learning and the formation of learning partner models retrievable afterwards. In this experimental study (N = 70), we investigated potential schema effects in partner modeling: Participants were first provided with descriptions of two learning partners (expert vs. novice of an area) and with their knowledge profiles consisting of knowledge levels (high vs. low) regarding certain topics of this area. In a memory test, participants had to remember the specific knowledge levels of both partners. Partner modeling was analyzed with multinomial processing tree models, as these models disentangle memory and guessing, which are often confounded when schema effects on memory are examined. High and low knowledge levels were equally well remembered for both partners. However, participants showed metacognitive biases, expecting their memory to be better for high knowledge levels. Additionally, knowledge profile estimations revealed that while novices’ knowledge was estimated accurately, experts’ knowledge was overestimated. We discuss the results and potential benefits of using multinomial processing tree models in the learning sciences for and beyond analyzing schema effects in partner modeling.

Keywords: Computer-supported collaborative learning, Group awareness tools, Partner modeling, Schemas, Multinomial processing tree modeling

1. Introduction

Collaborative learning is an omnipresent component of education in nearly all formal and informal settings. However, effective social learning relies on learners being aware of their learning partners’ knowledge profiles (i.e., what they know or do not know about a learning topic). If they are, targeted questions and explanations can be adapted to the learning partners and facilitate knowledge construction (Erkens & Bodemer, 2019; Webb, 1989). Such cognitive group awareness, i.e., the perception of competence or knowledge levels (Janssen & Bodemer, 2013), can be enhanced in computer-supported collaborative learning (CSCL) settings: Group awareness tools (GATs) collect, transform, and present relevant knowledge-related characteristics of learning partners (Bodemer et al., 2018) and learners can more effortlessly perceive and use relevant information about their partners’ knowledge. However, especially in (long-term) repeated collaboration, it is crucial to memorize knowledge-related information about learning partners: Does the partner know a lot about a topic and thus can be asked questions, or does the partner have low knowledge and might need explanations? Indeed, these specific partner modeling (PM) processes, i.e., estimating others’ knowledge (Dillenbourg et al., 2016), can have beneficial effects on learning (Sangin et al., 2011).

Heterogeneous groups (regarding expertise) sometimes outperform homogeneous groups because of useful interactions between learners, such as asking questions to an expert
and providing explanations to a novice (Webb, 1989). In some cases, a learner might learn collaboratively with both: A high knowledge partner (expert) and a low knowledge partner (novice) regarding an area (e.g., animals). Social schemas (e.g., “animal experts”) can help us navigate through the world, but they can also be dangerous as they can be inaccurate (Schnaubert & Bodemer, 2022). It is reasonable to assume that an expert has rather high knowledge in many different content-specific topics. However, it is also possible that the expert has low knowledge in some topics. Schemas carry the risk of overestimating partner competence and corresponding biases. In an experimental study, we explore potential schema effects and investigate whether PM processes differ for expert and novice partners. While, for example, Erkens and Bodemer (2019) found no differences in PM for high and low knowledge partners, we further disentangle PM processes and use analysis methods, i.e. multinomial processing tree (MPT) models, which are novel in CSCL research, but often used to analyze schema effects in source memory or person memory research.

1.1 Expected partner modeling advantage for schema-inconsistent information

Source memory and person memory research suggest an inconsistency-effect: Information which is inconsistent with a schema (here: partner expertise) is better remembered than schema-consistent information (e.g., Bell et al., 2015). For example, it is better remembered when trustworthy appearing persons cheat or untrustworthy appearing persons cooperate (Mieth et al., 2021). Here, we test whether such results can be transferred to PM processes when specific knowledge levels of experts and novices must be remembered. Biased modeling of learning partners’ knowledge levels could hinder efficient knowledge exchange processes in collaborative learning, such as asking questions about the right topics or giving explanations adjusted to partners’ low-level topics. Based on the inconsistency-effect, we expect that memory for low knowledge levels of an expert should be more accurate than memory for high knowledge levels (H1a) and memory for high knowledge levels of a novice should be more accurate than memory for low knowledge levels (H1b).

1.2 Expected metacognitive beliefs of higher schema-consistent modeling accuracy

Beyond actual memory processes, we explore participants’ metacognitive beliefs regarding their PM accuracy, as these beliefs themselves can influence behavior (Schaper & Bayen, 2023). In contrast to the inconsistency-effect observed in actual memory performance, when people judge their own memory performances, schema-consistent effects can be found: People assume their memory to be better for schema-consistent information than for schema-inconsistent information. For example, Mieth et al. (2021) have shown that while cooperation by untrustworthy people and cheating by trustworthy people is better remembered, people assume their memory to be better for the opposite. As these metacognitive illusions can lead to poor regulation and control processes, we considered metacognitive assumptions in addition to actual memory performances. Based on metacognitive illusions in source memory and person memory, we anticipate schema-consistent assumptions of learners: Participants’ metacognitive assumptions of their own PM accuracy for an expert should be higher for high knowledge than for low knowledge (H2a), and for a novice partner, assumptions of own PM accuracy should be higher for low knowledge than for high knowledge (H2b).

1.3 Exploratory analysis and study aim

We will also exploratively examine global PM accuracy: We tested the accuracy of perceived knowledge profiles regarding the amount of high and low knowledge topics of both partners. This study investigates potential schema effects in PM processes using MPT models, accurately assessing (schema driven) memory for different knowledge levels by disentangling memory and guessing processes (Bröder & Meiser, 2007). The novel use of MPT models in PM processes and combining paradigms used in different research fields can enrich CSCL research. This consequently enables a holistic view on long-term effects of tool usage in CSCL to better find instructions beneficial for learners and theory driven implications for tool design.
2. Methods

$N = 70$ participants took part in the experimental study. Some descriptive data and dependent variables have some missing data, resulting in different degrees of freedom. The age range ($N = 67$) was between 18 and 40 years ($M = 22.81, SD = 4.40$), and the sample was a student sample ($N = 69$, 94.20% students, 49 female, 20 male). We used a $2 \times 2$ within-subject design with the factors partner schema induced by providing expertise labels and descriptions (expert vs. novice) and partner knowledge level for specific learning topics (high vs. low).

The procedure is illustrated in Figure 1. Participants were first informed that two persons would be presented with information about their knowledge of certain animals. Also, participants would (allegedly) have to read texts about certain animals, learn collaboratively with the presented persons, and finally take a knowledge test. Next, descriptions about the partners (expert vs. novice) followed. Both were described as a 24-year-old man: While the expert was described as a zoology student with a passion for animals and who is seen as an expert regarding animals, the novice was described as a history student without pets, who is seen a novice regarding animals. Both partners’ presentation order was counterbalanced.

![Figure 1](image-url)

**Figure 1.** Schematic illustration of the key phases of the experiment.

In the partner model formation phase, participants were informed that both partners conducted a test and that their knowledge levels about specific animals would be presented (high vs. low). In a norming study, $N = 14$ participants rated 160 animals regarding their knowledge about them. We chose 80 animals with moderate ratings. From this pool, 20 animals were ascribed to the expert and 20 to the novice under different randomization conditions. On separate pages for each partner, a GAT presented knowledge profiles with knowledge levels about 20 animals in a table. Animals were displayed on the right side with their according knowledge level on the left side, with high levels shown in green boxes and low levels in white boxes (Figure 2a). Participants had 3 to 5 minutes to view the knowledge profile of each partner. While the expert’s knowledge profile consisted of 12 high and 8 low knowledge levels, the novice’s pattern was reversed to ensure credible schemas.

![Figure 2](image-url)

**Figure 2.** Schematic illustrations of the partner model formation (a) and partner model retrieval phases (b).

Participants next provided metacognitive assumptions of their perceived modeling accuracy. On scales from 0% to 100%, participants indicated for each of the 4 categories (expert high, expert low, novice high, novice low) for how many of the remembered topics the correct knowledge level (high, low) was memorized. After providing demographic information, the partner model retrieval phase started (Figure 2b). Here, for each partner the 20 animals of the partner model formation phase were presented intermixed with 20 new animals. On separate pages for each partner, participants had to remember the information of the GAT in the learning phase: Participants indicated whether each animal was associated with high knowledge, low knowledge, or was new. On the final page, participants provided knowledge profile estimations (global PM): For both partners, the proportion of high and low knowledge levels had to be indicated by sharing 100% between both categories (“Please indicate for what percentage of the animals the expert and novice had high and low knowledge levels”).
3. Results

3.1 Multinomial processing tree models: Schema effects on partner modeling (H1)

When memory for context information of items (here, high vs. low knowledge for certain topics) is assessed, especially in contexts with schemas (expert vs. novice partners), analyses based on hit and false-alarm rates can lead to the conclusion that schema-consistent information can be memorized better due to guessing biases (Bröder & Meiser, 2007). Multinomial processing tree (MPT) models offer a solution: Based on observed category frequencies (e.g., number of “high” responses to high knowledge topics), these models disentangle different memory and guessing processes and enable probability estimations of these processes (for an overview of MPT models, see Erdfelder et al., 2009). Here, we adapted the two-high threshold MPT model of source monitoring (Bayen et al., 1996) for our study purposes (Figure 3) to assess (schema effects on) PM in collaborative learning. The model illustrates that combinations of different cognitive processes can lead to the same response in the partner model retrieval phase. For example, the response “high” to an actual high knowledge topic may be based on correct memory processes ($D_{\text{high}} \times d_{\text{high}}$) or guessing processes, e.g., $D_{\text{high}} \times (1 - d_{\text{high}}) \times a$.

![Figure 3](image)

Figure 3. Multinomial processing tree model of source monitoring (Bayen et al., 1996), adapted for our study purposes with two sets (for the expert and novice). Rectangle on the left side represents the type of presented topic in the partner model retrieval phase. Index $i$ denotes the associated knowledge level regarding that topic, $i \in \{\text{high, low}\}$. Rectangles on the right side represent the possible answers of participants. Letters along the links represent probabilities of cognitive processes: $D = \text{Recognizing that a topic was old (i.e., presented before)}$; $d = \text{Remembering the associated knowledge level of an old topic (specific partner modeling)}$; $b = \text{Guessing that a topic was old}$; $a, g = \text{Guessing that a recognized (a) or unrecognized (g) topic was associated with a high level}$. The processing tree for new topics follows a similar principle (for an example, see Bayen et al., 1996).

To use our model, an identifiable base model with certain parameter restrictions must first be found. None of the identifiable base models (see Bayen et al., 1996, Figure 4) fit our data, models 5: $G^2(2) > 31.76, p < .001$, model 4: $G^2(4) = 277.18, p < .001$. We thus had to find a base model which fits the data with restrictions across both conditions. A model with the restrictions $D_{\text{Low}} = D_{\text{New}}$, $D_{\text{Low}} = D_{\text{New}}$, $d_{\text{High, Expert}} = d_{\text{Low, Expert}}$, $d_{\text{Low, Novice}}$ shows good fit, $G^2(1) = 1.19, p = .276$. The restriction of the PM-parameters ($d = .76, SE = 0.03$) inherently means that specific PM does not differ within and between both partners: High and low knowledge of experts and novices seem to be equally well modelled. Consequently, the hypotheses $H1a$ and $H1b$ cannot be supported by the data: Our findings do not provide evidence for an inconsistency effect in PM. Participants demonstrated relatively accurate PM abilities with a 76% probability of remembering the knowledge levels.

3.2 Schema effects on metacognitive assumptions of partner modeling accuracy (H2)

A 2 × 2 repeated-measures ANOVA (descriptive data in Figure 4a) revealed no significant main effect of partner schema on metacognitive assumptions, $F(1, 67) = 0.85, p = .360, \eta^2 = .01$. There was, however, a significant main effect of knowledge level, $F(1, 67) = 56.74, p < .001, \eta^2 = .46$. Although actual memory performances between high and low knowledge levels did not differ (see section 3.1), participants assumed that they remembered high knowledge better than low knowledge. The ANOVA further revealed a significant interaction effect, $F(1, 67) = 13.32, p < .001, \eta^2 = .17$ (all follow-up pairwise comparisons were significant, $t(67) > 2.78, p_{\text{holm}} < .009$): Participants assumed better memory for high knowledge compared to low knowledge both for the expert (contradicting $H2a$) and the novice (supporting $H2b$), but the difference was greater for novice partners than for expert partners.
3.3 Global partner modeling accuracy: Perceived knowledge profiles (explorative)

We further tested global PM accuracy (Figure 4b). One sample \(t\)-tests revealed that the expert’s knowledge was overestimated: Mean estimations of the number of topics with high knowledge were higher than 60\%, \(t(65) = 3.08, p = .003, d = 0.38\). However, the proportion of high and low knowledge topics of the novice was accurately estimated, \(t(65) = 1.00, p = .319, d = 0.12\). The novice’s knowledge was only descriptively overestimated.

4. Discussion

Group awareness tools (GATs) provide knowledge related information about learning partners to support effective CSCL processes. Sometimes, knowledge related information needs to be remembered in retrospect (partner modeling, PM), both in short- and long-term collaborations. We investigated whether learners model knowledge of experts and novices differently, predicting a memory advantage for unexpected information (e.g., knowledge gaps of experts). However, we could not observe such an inconsistency-effect in PM: High and low knowledge were equally well encoded for expert and novice partners (contradicting \(H1\)). For educational practice, this implies that in short-term collaboration—for example in heterogeneous groups with high and low knowledge students—such PM biases do not impair collaborative learning processes but allow for asking and answering questions properly adapted to the learning partners. However, other studies found the inconsistency effect. We can thus assume that there may be conditions and moderating variables affecting the results. For example, Ehrenberg and Klauer (2005) found the inconsistency-effect primarily in contexts with higher cognitive load during encoding and longer retention-intervals, which are given in many educational settings. Consequently, manipulations of collaboration duration could be employed to test whether the inconsistency effect is generally absent in PM, or whether our results are limited to short-term collaborative learning situations.

We further considered metacognitive assumptions, as wrong metacognitive assumptions can also influence control processes and study behavior (Schaper et al., 2023). Here, learners generally assumed memory for high knowledge to be better than for low knowledge (supporting \(H2a\), but contradicting \(H2b\)). When students mistakenly assume that they would remember high partner knowledge better than low knowledge, they might ask only about their partners’ knowledge gaps, even though asking for high knowledge topics would give them the opportunity to immediately access the best explanations. Metacognitive prompts can enhance students’ metacognitive awareness when collaborating (Teng, 2022). Integrating prompts into GATs can assist students recognizing the metacognitive illusion of alleged better modeling of high levels, thereby countering the illusion and fostering better help-seeking.

Exploratively examining knowledge profile estimations (global PM) revealed that novices’ knowledge was estimated accurately, which is desirable, as both, over- and underestimations of novices by more knowledgeable partners can impair novices’ learning (Wittwer et al., 2008). Experts’ knowledge, however, was overestimated, which might pose a
danger: Students might over-trust experts and their explanations, without adequately verifying their produced learning contents. Again, metacognitive prompts or cues—delivered through guidance from educators or integrated into GATs—could potentially make learners aware of potential overestimations and aid them in critically questioning their assessments.

From a methodological perspective, MPT models can disentangle memory and guessing processes and thus offer diverse applications in the field of learning sciences. They have the potential to reveal cognitive biases beyond (potential) schema effects in PM: For example, learners might project their own knowledge onto peers when failing to remember their knowledge (Nickerson, 1999). MPT models could test this assumption by assessing whether the guessing parameters of the model differ based on students’ prior knowledge. Accurately disentangling different PM processes and biases can help to derive sound implications for instructors and tool designers. While this study was a first attempt to use MPT models to test the impact of collaboration tools on PM processes, further drawing on methods, paradigms, and findings of research areas such as source memory holds the potential to enrich CSCL research, ultimately benefiting learners.

Acknowledgements

The authors thank Jamie-Lee Tripke for data collection in the course of her Bachelor's Thesis.

References


Exploring Group Formation Strategies in Computer-Supported Collaborative Learning: A Systematic Review

Jiamin TANG, Huihan ZHOU, Yajing TAN & Guang CHEN*
Faculty of Education, Beijing Normal University, China
*guang@bnu.edu.cn

Abstract: In this study, three independent reviewers conducted a systematic review of articles on group formation strategies in computer-supported collaborative learning (CSCL) and synthesized and structuralized the involved grouping attributes and evaluation indicators with some other features. The study aims to contribute to efficient and effective group formation in educational practices in CSCL environments.

Keywords: Group formation strategies, computer-supported collaborative learning, systematic review

1. Introduction

With the proliferation of information and communication technologies (ICTs) and the far-reaching pandemic, CSCL has further developed in education, whose benefits when appropriately applied have been validated by substantial studies (Costaguta, 2015).

The group composition affects the learning outcomes and the group development lifecycle, which are crucial in fostering an effective and efficient CSCL setting (Krouska et al., 2019). Thus, exploring effective grouping strategies in the CSCL environment and their impact on CSCL becomes practically necessary.

Previous studies have focused on 2004-2018 articles about grouping strategies, while literature reviews evaluating the effects of those strategies are insufficient. Furthermore, although online learning has been integrated into pedagogical practices with unprecedented depth and scope since 2019 due to the pandemic, the group formation strategies in CSCL from 2019-2023 have not been deeply examined. Therefore, to fill the gaps and compare trends before and after the pandemic, the researchers searched articles on group formation strategies in CSCL in the past 12 years. Due to the time limitation and the concerns for comprehensiveness and credibility, Web of Science was chosen as the sole database for search. Hence, our research questions are as follows:

Q1: What group formation strategies have been adopted in the CSCL environment?
Q2: What has been the influence of those strategies on CSCL?

2. Method

In this study, the English-language journal articles or conference articles of group formation strategy in Computer-Supported Collaborative Learning that published from 2012 to 2023 were searched in the Web of Science database, and The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework and snowballing were the primary guide for this study. The three reviewers worked collectively to discuss the criteria of articles’ inclusion and exclusion and independently coded articles, reaching a high level of consistency (Kendall’s W = 0.81), and 17 articles were finally included. For the 17 final included articles, this study analyzed the basic information, counted the effect of group formation strategy, and used Theme analysis method analyzed the group formation strategy (Virginia et al., 2006).
3. Result and Discussion

3.1 Descriptive Information of the Literature

As for the time distribution (Fig 1. a), the number of published articles remained stable from 2012 to 2019 and showed a growing trend from 2019 to 2022. As for the country distribution (Fig 1. b), Chinese researchers paid more attention to research in this field. In terms of the distribution of the education stage (Fig 1. c), the research on group formation strategies in CSCL was mainly focused on higher education (82%). As for the subject distribution (Fig 1. d), group formation strategies in CSCL mainly focused on STEM disciplines (72%). As for the sample size, the number of participants presented a state of polarization, with the number of participants less than 50 and more than 200 accounting for the majority.

![Figure 1. The descriptive information graphs of the literatures](image)

3.2 Group Formation Strategy

The initial codes had been distilled from the group formation strategies of the included articles and combined into an initial thematic web map (shown in Figure 2. a), describing the involved strategies from eight thematic dimensions, such as "Student Learning Performance" or "Personality." These themes are interdependent with links between them, such as "Spatial Competence" in "Intelligence" and "Gender" in "Demography Variable" (Zhong et al., 2023). Although differences and commonalities coexist in these eight themes, they were all retained temporarily before the sequent reviews.

![Figure 2. The thematic web map](image)

Next, the individual themes and the corresponding codes under the themes were verified and proofread, which formed Figure 2. b. The researcher incorporated “Personality” and "Intelligence" into the theme "Intrinsic Attributes of Students"; while “Learning Interest” and "Learning Style" and “Learning Performance” were included in the theme "Learning Variables". In addition, those highly similar codes, such as “Academic Performance” and “Learning Performance”, were merged into “Student Learning Performance”. The final thematic network diagram was developed after the themes’ definitions and the ultimate optimization (Figure 2. c). The researchers found that those strategies could be divided into explicit attributes and implicit attributes according to the difficulty of their measurement. Explicit attributes are easy to obtain and measure, while implicit attributes...
require specific measuring tools to obtain results which may have more options and may change with different measuring tools. Additionally, the demography variable belonging to explicit attributes can be mixed with the learning variable or the intrinsic attributes students belonging to implicit attributes together as group formation strategy rather than just alone.

3.3 The Evaluation of Group Formation Strategy

Most of the studies evaluated group formation strategies with different involved attributes using a mixture of indicators, some researchers reaching a consensus conclusion on the impact of the same group formation attributes, and some not. For example, As for the influence of gender as a group formation attribute, the researchers reached a consensus conclusion on the role of gender composition in groups regarding its influence on learners' attitudes. In addition, no consensus conclusion has been reached on the influence of learning style and achievement level as grouping attributes, which requires further research. Detailed information can be obtained from figure 3.

<table>
<thead>
<tr>
<th>Group formation attributes</th>
<th>Gender</th>
<th>Role</th>
<th>Group size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation indicators</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conclusion</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. The Involved Attributes of Group Formation Strategies and Evaluation Indicators

4. Conclusion

Based on a review of group formation strategies in the context of computer-supported collaborative learning from 2012 to 2023, this study constructed a theme map displaying the classification of the attributes involved in group formation strategies, analyzed the influence of the group formation strategies. The limitation of this study is that only one database was consulted for the data source of this study. Therefore, the data source should be expanded to include more literature to provide a broader and more holistic perspective in future studies.

References

A Comparative Analysis on the Effects of Cognitive Tools in Data Inquiry Cultivation

Hui ZHANG*, Bian WU, Yi-Ling HU & Yu-Jie XU
Department of Education Information Technology, East China normal University, China
*51214108043@stu.ecnu.edu.cn

Abstract: The advent of the big data era has elevated the importance of understanding, analyzing, and mining data, necessitating data science education with cognitive tools aiding data inquiry. However, limited appraisals exist, especially regarding diverse tools’ impact on data inquiry cognitive aspects. This study addresses this void by examining cognitive tool impact on data inquiry. The study collected discourse and questionnaire data generated during online collaborative tasks. Epistemic network analysis and difference testing unveiled cognitive pattern disparities, learning motivation, cognitive load, and self-efficacy variations between groups. Block-based group displayed robust cognitive connections in data understanding and preparation. Text-based group focused on modeling and optimization iterations. Motivation and load didn’t differ significantly, yet block-based group showed higher self-efficacy. Study concludes by acknowledging limitations and suggesting future research directions.

Keywords: Data inquiry, cognitive tools, CSCL, human-computer interaction

1. Introduction

Data inquiry is the process of extracting meaningful and valuable information from data, including the identification of understandable patterns and relationships, as well as the construction of representative data models. In recent years, many researchers have provided various perspectives on the connotative structure of data inquiry. Most of them perceive data inquiry as a collection of data skills, encompassing elements such as data collection, pre-processing, analysis, model building, and evaluation (Donoho 2017).

Various cognitive tools have emerged to support data practice and learning. These tools can be categorized into three types: block-based, text-based programming, and menu-selected (Bart et al. 2020). Block-based tools, such as RapidMiner and IBM SPSS Modeler, offer encapsulated modular systems and user-friendly graphical interfaces, enabling quick prototyping and validation of predictive models. They simplify the entire analysis process and provide instant gratification. On the other hand, text-based programming tools like RStudio and Python offer greater flexibility and scalability. They allow users to explore algorithmic details, create customized program algorithms, and offer system openness.

Block-based cognitive tools offer a potential resolution to the educational challenge of data inquiry. Their algorithmic intricacies are concealed, enabling learners to concentrate on configuring and adapting predefined algorithms and parameters. By seamlessly connecting modules through drag-and-drop actions, learners swiftly execute data inquiry, with a shallow learning curve fostering confidence and problem-solving belief. Modular encapsulation delegates computational tasks to cognitive tools, enabling learners to focus on problem solving, expanding mental capacity and reducing cognitive load (Javadpour 2022). This phenomenon is well-established in computational thinking development. Additionally, graphical cognitive tools deliver instant interactive feedback via module interconnections, prompting immediate data-inquiry knowledge reflection, potentially heightening motivation to learn.

However, there have been no studies exploring the differences in the effects of block-based and traditional text-based cognitive tools on learners' data inquiry abilities. At the same time, researchers have pointed out that future studies should explore whether certain data science practices are influenced by tool affordance (Jiang and Kahn 2020). In order to better
design teaching and learning based on block-based data science cognitive tools, clarify how cognitive tools affect learning processes and knowledge construction, this study proposes the following research questions:

1) What are the differences in the frequency distribution of data inquiry cognitive elements between the block-based and the text-based tools in online collaboration activities?

2) What are the differences in the cognitive patterns of data inquiry between the block-based and the text-based tools in online collaboration activities?

3) Are there any differences in learning motivation, cognitive load, and self-efficacy between the block-based and the text-based tools?

2. Method

2.1 Research Context and Participants

This empirical study was conducted during the Fall Semester of 2022-2023 at a comprehensive university in eastern China, focusing on the "Educational Data Mining" course. Designed for senior college students pursuing careers in education, the study participants were selected from two classes within the same university department, totaling 44 individuals - 10 males and 34 females. All participants had not received prior data mining instruction. Collaborative learning was facilitated in groups of 3-4 learners. One class (16 students, 5 groups, average age 21) used text-based cognitive tools, while the other class (28 students, 9 groups, average age 23) employed block-based cognitive tools.

2.2 Research Process

The experimental procedure is illustrated in Figure 1. The study spanned 18 weeks, comprising a 14-week instructional phase dedicated to learners' acquisition of data inquiry skills. Both classes were taught exactly the same by a veteran instructor with over a decade of teaching experience. Subsequently, participants undertook data analysis projects during the final 4 weeks. Upon course completion, all students were required to complete questionnaires assessing cognitive load, learning motivation, self-efficacy, and satisfaction. The questionnaire completion process typically lasted around 10 minutes.

![Figure 1. The flowchart of research experiment.](image)

The two student cohorts employed distinct cognitive tools – block-based and text-based – for their data inquiry tasks. The block-based group utilized the RapidMiner platform, a robust data mining tool characterized by its visual workflow interface, facilitating intuitive design, execution, and assessment of diverse data mining tasks. Conversely, the text-based group leveraged the RStudio platform, an interactive development environment based on the R programming language. This platform encompasses a variety of functions, including a code editor and data viewer, facilitating efficient organization and management of code and data.
2.3 Data Collection

In the project’s final 4 weeks, each group conducted 5 online discussions lasting 1 to 3 hours, recorded for analysis. The study employed adapted percentage scored questionnaires for learning motivation (6 items, Cronbach’s $\alpha = 0.848$), cognitive load (2 items, Cronbach’s $\alpha = 0.836$), and self-efficacy (3 items, Cronbach’s $\alpha = 0.843$). Learning motivation drew from Lau & Lee (2008), measuring knowledge mastery and grades on a five-point scale. Cognitive load, adapted from Leppink et al. (2013), assessed internal and external load using a nine-point scale. Self-efficacy, from Tsai et al. (2020), focused on data inquiry self-efficacy with a five-point Likert scale.

2.4 Code Framework

A coding framework was applied to analyze cognitive elements in data inquiry for frequency and patterns. Adapted from Wirth & Hipp (2000), a recognized data mining model, the initial framework had 6 codes. Grounded theory guided manual analysis of pre-experiment conversation data, refining the framework. The final version contained 7 codes: planning, understanding, preparation, modeling, evaluation, technical inquiries, and tangential topics. For code breakdown, meanings, and case examples, consult Table 1.

To ensure coding framework reliability, two coders independently processed a randomly selected 15% of session data, defining content analysis units as uninterrupted student presentations. Results showed substantial agreement, Kappa score at 0.784. Coders discussed discrepancies to enhance shared understanding and scheme interpretation. Remaining data (8,615 sentences) was then independently coded, resulting in 5,986 coded data points.

Table 1. Collaborative Conversation Data Coding Framework

<table>
<thead>
<tr>
<th>Code</th>
<th>Meaning</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan</td>
<td>Develop plans to achieve the goals. Review models and processes, issues that arise, and identify areas for improvement in subsequent work.</td>
<td>For the next step, we may have to continue to extract the features. I'm missing a descriptive statistic for each type of question, such as the total value of the mcss type and its average value, and time.</td>
</tr>
<tr>
<td>Understand</td>
<td>Understand project goals and requirements, understand and infer data context and types, hidden trends, anomalies, patterns in data, explore data distribution, correlations and trends.</td>
<td>One is to determine whether he answered the question or not, and the other is to calculate the threshold value. These are the two problems that need to be solved in order to go ahead and make predictions based on these.</td>
</tr>
<tr>
<td>Preparation</td>
<td>Fixing data problems, converting data formats, merging multiple data sets, extracting new features from data, transforming features, selecting some data rows and features.</td>
<td>We need to standardize the test dataset. I roughly sieve a few features, you guys look at. We can count the type of operation. We can also add other columns, such as how many times to enter, how many times to exit, how many times to draw.</td>
</tr>
<tr>
<td>Modeling</td>
<td>Find, search, introduce, and filter algorithmic models that can be used, build models using cognitive tools, understand the meaning of model parameters, and change parameters to improve model performance.</td>
<td>Can you use generalized linear regression, but we're not a continuum here. There should be a template from when you wrote the logistic regression before. Take it and change it. I adjusted the tree model, because the previous 1500 may overfitting, so I set it to 1000, may be a little better.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Select evaluation metrics to assess model performance, calculate selected metrics using software manipulation or code programming, and understand</td>
<td>I see logistic regression can convert probabilistic output to labeled output, it can also output precision confusion matrix, exact value. How to use kappa for logistic regression? AUC is still 0.3536; Kappa is still 0, very low.</td>
</tr>
</tbody>
</table>
2.5 Data Analysis

Technology problem

Only address technical issues that arise during interaction with RStudio and RapidMiner.

Irrelevant topic

Distractions, or project related emotional expressions.

If kappa is equal to 0, it means that the two judgment results are caused by chance.
No matter how I access it, I cannot access its core content. So, I'm stuck here, stuck for a long time, stuck me for almost four or five hours. R language is too disgusting. I'm infected COVID-19 too.

2. Results

3.1 What Are the Differences in the Frequency Distribution of Data Inquiry Cognitive Elements Between the Block-Based and the Text-Based Tools?

A total of 3062 codes were generated in the Block-based group, and 3148 codes were generated in the text-based group. Specifically, the block-based group had a significantly higher cognitive frequency than the text-based group in planning ($\chi^2 = 11.239, p < 0.01$) and understand ($\chi^2 = 164.512, p < 0.01$). Code frequencies for modeling ($\chi^2 = 102.044, p < 0.01$), evaluation ($\chi^2 = 88.256, p < 0.01$), and irrelevant topics ($\chi^2 = 83.195, p < 0.01$) were significantly higher in the text-based group. See Table 2 for more detailed results.

<table>
<thead>
<tr>
<th>Code</th>
<th>block-based</th>
<th>text-based</th>
<th>$\chi^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan</td>
<td>274</td>
<td>190</td>
<td>6.04%</td>
<td>11.239</td>
</tr>
<tr>
<td>Understand</td>
<td>652</td>
<td>259</td>
<td>8.23%</td>
<td>164.512</td>
</tr>
<tr>
<td>Preparation</td>
<td>1239</td>
<td>1160</td>
<td>36.85%</td>
<td>0.034</td>
</tr>
<tr>
<td>Modeling</td>
<td>268</td>
<td>514</td>
<td>16.33%</td>
<td>102.044</td>
</tr>
<tr>
<td>Evaluation</td>
<td>273</td>
<td>499</td>
<td>15.85%</td>
<td>88.256</td>
</tr>
<tr>
<td>Technology problem</td>
<td>147</td>
<td>117</td>
<td>3.72%</td>
<td>1.92</td>
</tr>
<tr>
<td>Irrelevant problem</td>
<td>209</td>
<td>409</td>
<td>12.99%</td>
<td>83.195</td>
</tr>
<tr>
<td>Total</td>
<td>3062</td>
<td>3148</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

3.2 What Are the Differences in the Cognitive Patterns of Data Inquiry Between the Block-Based and the Text-Based Tools?

The connection lines between the nodes represent the relationship between the data inquiry elements, and the thickness of the connection line reflects the number of times the two nodes appear simultaneously in the students' utterances. By observing the ENA diagrams and comparison plot of two groups, as shown in Figure 2, we found that the block-based group's plan-understand, understand-preparation are highly correlated. In the text-based group, evaluation frequently co-occurs with preparation, modeling, and technical problem. In addition to this, preparation is also more highly relevant to modeling and technical problem.
3.3 Are There Any Differences in Learning Motivation, Cognitive Load, and Self-Efficacy Between the Block-Based and the Text-Based Tools?

The Block-based group exhibited a mean learning motivation of 76.000, while the text-based group had 72.667. T-test results ($t = 0.814$, $p = 0.427 > 0.05$) indicated no significant difference between them. Comparing cognitive load, the block-based group averaged 81.482, and the text-based, 83.704. Independent sample t-test ($t = -0.787$, $p = 0.438 > 0.05$) revealed both groups had higher cognitive load, but insignificantly different. However, self-efficacy showed significance, $U = 66.000$, $p = 0.047 < 0.05$, with block-based group at 72.889, and text-based group at 63.111 mean. See Table 3 for detailed results.

4. Discussion and Conclusion

This study extensively examined how distinct cognitive tools influence students' data inquiry skills by analyzing data inquiry patterns within block-based and text-based groups. Both tool types exhibit distinct merits in enhancing data comprehension and modeling.

4.1 Block-Based Cognitive Tools Make It Easier for Beginners to Enter Data Inquiry

The block-based group exhibited significantly higher cognitive frequencies for planning and understanding than the text-based group. Furthermore, ENA network connections, particularly between understanding and preparation, were markedly stronger in the block-based group. Upon comprehending task objectives and data meanings, the block-based group strategized converting complex datasets into a streamlined format using the pivot module operator in RapidMiner. This transformation enhanced data comprehension, revealing trends like incomplete question responses, which the original dataset obscured. As understanding deepened, the group refined their inquiry plans. This iterative process, underpinned by solid data contextualization, facilitates pattern identification (Wilkerson, Lanouette, and Shareff 2021), emphasizing data inquiry’s importance. Conversely, the text-based group encountered coding challenges, causing disruption between planning, understanding, and preparation. This resulted in decreased self-efficacy. Similar findings were noted by Price and Barnes (2015) in computational thinking comparisons. Novices, facing difficulties, may cease autonomous learning after few attempts (Tawfik, Payne, and Olney 2022).
Text-based group exhibited significantly higher frequencies of modeling and evaluation than block-based group. In-depth analysis revealed text-based group's model selection and parameter challenges, necessitating external resource usage (blogs, papers). This process enhanced their model understanding and internal models, aiding subsequent feature screening goals, as supported by Oh & Oh (2011). In contrast, block-based group relied on tool-provided modules within its interface, missing external learning opportunities. Although attempting more models, they lacked direction and understanding. Instructors must ensure resource exploration beyond tool for less reliance on blind attempts.

Cognitive load and learning motivation didn't significantly differ. However, working memory constraints and simultaneous module identification hinder effective attention allocation to modeling and optimization. Text-based group efficiently treats multiline text codes, reducing element interactivity (Chen, Kalyuga, and Sweller 2017) and cognitive load. Encapsulating multiple modules into one may ease resource consumption.

Acknowledgements

This paper was supported by the East China Normal University graduate student international conference special fund.

References

Tsai, Chia-Lin, Moon-Heum Cho, Rose Marra, and Demei Shen. 2020. “The Self-Efficacy Questionnaire for Online Learning (SeQoL).” Distance Education 41(4): 472–89.
Argumentative Knowledge Construction and Certainty Navigation: A Comparison between Individual and Group Work

Wenli Chen\textsuperscript{a}, Eng Eng Ng\textsuperscript{a, b}, Su Guo\textsuperscript{a}, Su Junzhu\textsuperscript{a}, Xinyi Li\textsuperscript{b}, Siew Cheng Aileen Chai\textsuperscript{a} and Qianru Lyu\textsuperscript{a}

\textsuperscript{a}National Institute of Education, Nanyang Technological University, Singapore
\textsuperscript{b}Beijing Normal University, China
\textsuperscript{*}caryn.ng@nie.edu.sg

Abstract: This study investigated the extent to which levels of certainty impacted the argumentative knowledge construction in individual work and group work. Argumentative knowledge construction has been characterized into simple claims, grounds, qualifiers, counterarguments, and integrated replies to illustrate the components of argumentation and nature of resolving conflicts in argumentation where certainty levels have been divided into uncertain, neutral, and certain. Findings showed that individual and group work differed significantly in terms of levels of certainty for simple arguments and counterarguments. Study implications were discussed.

Keywords: Knowledge construction, collaborative argumentation, individual writing, group work, argumentative writing

1. Introduction and Literature Review

In Computer-Supported Collaborative Argumentation (CSCA) research, the argumentative knowledge construction model seeks to understand how social interactions and argumentation processes contribute to knowledge construction and learning (Nguyen & Young, 2022). Through engaging in social interactions and argumentation with others, individuals actively navigate uncertainty by constructing their understanding of a topic, negotiating meaning, and refining their ideas. Existing research on the area of knowledge construction has thus far centered on facilitating knowledge construction through placing the individual in the context of the community around him or her and building on discourse moves and reasoning to construct arguments (Nguyen & Young, 2022). Usage of words indicating certainty has been an area of focus in social interactions as they navigate between advancing their own knowledge and resolving conflicts with others (Tabak & Baumgartner, 2004; Tan et al., 2016). Uncertainty in argumentation in navigating conflict resolution is also demonstrated by using more hedging words (“possibly”, “maybe”, and “potentially” etc.) to question and evaluate each other’s claims and evidence (Martinez-Maldonado et al., 2013). However, hedges such as “may” or “perhaps” show a tentativeness to commit i.e., uncertainty. Word such as “actually”, showed stronger commitment. Our study investigated the extent to which levels of certainty used would impact argumentative knowledge construction when comparing individual work and group work.

In individual writing, students engage with deep learning of the content, clarify their thoughts and ideas, identify gaps in understanding, develop soft skills, and explore new perspectives on a topic (Harney et al., 2017; Latifi et al., 2020, 2021; Noroozi et al., 2022; Storch, 2005; Valero Haro et al., 2020; Weinberger & Fischer, 2006). Collaborative writing happens when group members contribute their perspectives and ideas to the group’s collective efforts, working together to achieve learning outcomes while honing technological literacy, problem-solving, critical thinking, innovation, shared responsibility, belongingness within the group, better mastery of the concepts, interrelationships among concepts, and integration of new concepts with prior knowledge (Asterhan & Schwarz, 2007; (Clark & Sampson, 2007; Mayweg-Paus et al., 2021; Roberts, 2004).
Argumentation fosters higher-order cognitive abilities as students dissent, evaluate, expand, and synergize claims, evidence, and counterarguments to reach conclusions (Harney et al., 2017; Latifi et al., 2020, 2021; Valero Haro et al., 2020). Such higher-order thinking skills involve knowledge construction, concept integration, critical thinking, and reflective judgments (Harney et al., 2017; King, 2002; Lu & Zhang, 2013). Students move from individual work/argumentation, preparing themselves from individually constructing their arguments to collaborative argumentation. Individuals move from personal viewpoints to resolving conflicts and converging at a joint and shared knowledge in argumentation. Through social interactions (Hou & Wu, 2011), students co-construct knowledge as interpersonal interactions connect academic discussions and argumentative task coordination (Hou & Wu, 2011).

The model for managing uncertainty developed by Chen et al., (2019) postulated that learners go through the first stage of (1) raising uncertainty, where they ask questions to clarify, critique, and challenge existing gaps in individual knowledge about a topic before moving on to the next stage of (2) maintaining uncertainty where learners seek to find solutions and evidence along with garnering varied perspectives (Nguyen & Young, 2022). Finally, learners reduce uncertainty by reconciling current knowledge with prior knowledge – integrating and resolving conflicts and gaps in knowledge. Although argumentation has been linked with uncertainty navigation in past studies, previous work has been focused on how uncertainty plays a tentative and suitable role in challenging and resolving conflicting arguments in argumentative discourse (Nguyen & Young, 2022).

Therefore, this study investigated the following research question and hypothesis:

Research Question: To what extent does level of hedging indicating certainty levels impact argumentative knowledge construction in individual work and group work?

Hypothesis: There was a significant difference in uncertainty using hedging for individual and group work in argumentative knowledge construction.

2. Method

2.1 Study Design and Participants

18 Singapore Secondary grade 3 students (15 years old; 10 males; 8 females) participated in this study. The teacher assigned students randomly to a total of 10 groups, with each group consisting of four students. They discussed collaboratively on an online collaborative argumentation platform known as AppleTree. The students in the class knew one another and experienced asynchronous group learning in other lessons. Students addressed the extent to which they agreed and provided evidence for their claims and rebuttals/counterarguments on a topic surrounding social issues: “Volunteering causes more harm than good. How far do you agree?” In the individual work/ideation phase (15 minutes), students individually constructed arguments on the given topic. When students moved into group synergy/work (20 minutes), they collaborate with one another to achieve a group consensus. Ethics approval and a signed written informed consent was obtained from all participants before data collection in April 2023.

2.2 Coding Argumentation of Artefacts

Content analysis was used to address our research questions and hypotheses to examine the certainty levels from students’ individual and group argumentative work. Argumentative knowledge construction is operationalized as the average scores of these following coded elements (see Table 1 below). Dichotomous coding was used. The presence of an element below was coded as 1 whereas 0 was given if the element was absent.

<table>
<thead>
<tr>
<th>Table 1. The Coding Scheme for Argumentation Knowledge Construction in Individual and Group Work (Weinberger &amp; Fischer, 2006; Weinberger et al., 2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Simple Claim</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Ground (with warrants)</td>
</tr>
<tr>
<td>Qualifier</td>
</tr>
<tr>
<td>Counter-argument</td>
</tr>
</tbody>
</table>

“I agree that volunteering causes more harm than good.”

“Yes, I agree because the person hosting this event wants to try their best to support the volunteers rather than those needing help.”

“However, sometimes when you volunteer to help someone the help you give may not really benefit them.”

“Volunteering spreads kindness around our neighborhood, precinct or even to further people, which could be a potential for more people being helped by volunteers to being volunteers to help those in need. However, some people feel that volunteering also causes negative impacts, such as people volunteering for financial gains and volunteering for work resume experiences to be written on. Some volunteering services are ineffective in helping people which causes more harm to the people in need than good.”

“Research also shows that short-term orphanage visits can cause damage to children’s development and emotional well-being, creating unhealthy short-lived attachments and separation anxiety. Children need constant sources of love and support.”

Level of certainty is indicated by the number of uncertainty and certainty words used in an argument (Nguyen & Young, 2022). Certainty and uncertainty words used by the learner were coded as 1 and then tabulated to compute a total score. If the participant used the same number of certainty and uncertainty words, the level of certainty would be coded as neutral (Nguyen & Young, 2022). Examples of uncertainty words are hedges such as “can” and “may be”, while certainty words include, “actually” and “definitely” (Nguyen & Young, 2022).

2.3 Procedure

Students went through individual work followed by group work and then back again, based on the Spiral Model of Collaborative Knowledge Improvement (SMCKI) (Chen et al., 2021): (1) Individual ideation: students work on their own to indicate the extent to which they agreed or disagreed with the given question with their supporting evidence; (2) Group synergy: individual claims are merged together in a single claim; (3) Refinement: students reflected on the feedback given to them by the other group, choosing to make changes if needed; and (4) Individual achievement: students write their own individual reflections after the whole argumentation activity (Chen et al., 2021). For this study, we examined students’ Individual Ideation work, and compared their individual work with their group work (Group Synergy).

The second and fourth authors coded students’ individual writing worksheets and argumentation graphs (see Figure 1) using the coding scheme (see Table 1). These findings indicated substantial levels of agreement between the independent coders (Kappa: 82.5%). We used NVivo version 12 and Microsoft Excel for the manual qualitative content analyses coding before proceeding to analyze our quantitative data in IBM SPSS version 28.

Figure 1. Sample of an argumentation diagram.
3. Results, Discussion, and Conclusion

This study investigated the extent to which differences were observed in students’ level of certainty as shown in hedging used in the context of individual and group social knowledge argumentation construction. The variables did not satisfy normality assumptions – potentially due to a small sample size. Using Spearman rho correlations (Table 2), significant associations ($p < .01$) were found between individual work qualifier and group work hedging used, group work ground and group work simple argument, group work ground and group work integrated replies, and group work ground and group work hedging used. It is expected that students’ certainty levels would decrease significantly as they move from individual work to group work. This is in line with past research which encouraged students to use hedging to express an openness to consider conflicting arguments against their initial position (Nguyen & Young, 2022). Nevertheless, students’ individual and group work were non-significant across other categories of argumentation. Thus, students’ certainty navigation remained consistent throughout argumentative knowledge construction in individual or group contexts.

Findings from one-way ANOVA (Table 3) revealed significant differences ($p < .05$) in certainty levels when forming simple arguments during individual work and counterarguments at group work. Post-hoc Bonferroni’s correction revealed significant differences between Certain and Neutral levels ($p < .05$, Mean Difference = .75; Standard Error = .26) when forming simple arguments during individual work, and between Uncertain and Neutral levels ($p < .05$, Mean Difference = .67; Standard Error = .18) when forming counterarguments at group work. Navigating uncertainty self-regulatory skills and attitudes requires time to hone and build (Xu et al., 2015). Although students are encouraged to express certainty, reconciling varied perspectives in argumentation required learners to take a tentative approach, hence, using uncertainty words (Nguyen & Young, 2022). It is possible that learners only become more aware of differences in synthesizing and expressing their arguments after moving on to group work from individual work. These findings added support that students did not merely copy and paste their individual work to group work, but rather, modified their work based on their learning experiences (Tsouvaltzi et al., 2017). Future interventions may focus on supporting group interactions and scaffolding students’ argumentative work to enable smooth online collaboration, familiarity with the challenges of communicating asynchronously (Oren et al., 2002), unshared knowledge in individual work translated into group work, levels of certainty impacted by vicarious learning via influence from other group members (Fischer & Mandl, 2001; Weinberger & Fischer, 2006), and emotional self-control (Nguyen & Young, 2022).

Table 2. Spearman rho Findings on Certainty and Argumentative Knowledge Construction Reflected in Individual and Group Work; IW: Individual Work; GW: Group Work; SA: Simple Argument; Q: Qualifier; CA: Counterargument; IR: Integrated Replies; HU: Hedging Used (Uncertain = 1, Neutral = 2, and Certain = 3)

<table>
<thead>
<tr>
<th>Variables</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>IW-SA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IW-Q</td>
<td>.28</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IW-CA</td>
<td>.22</td>
<td>.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IW-IR</td>
<td>.90</td>
<td>.63</td>
<td>.81</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-SA</td>
<td>.88</td>
<td>.59</td>
<td>.57</td>
<td>.69</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-G</td>
<td>.73</td>
<td>.20</td>
<td>.44</td>
<td>.63</td>
<td>.01**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-Q</td>
<td>.49</td>
<td>.75</td>
<td>.88</td>
<td>.45</td>
<td>.49</td>
<td>.89</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-CA</td>
<td>.50</td>
<td>.96</td>
<td>.07</td>
<td>.90</td>
<td>.06</td>
<td>.73</td>
<td>.49</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-IR</td>
<td>.63</td>
<td>.47</td>
<td>.38</td>
<td>.21</td>
<td>.06</td>
<td>.01**</td>
<td>.39</td>
<td>.63</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IW-HU</td>
<td>.85</td>
<td>.07</td>
<td>.64</td>
<td>1.00</td>
<td>.53</td>
<td>.91</td>
<td>.37</td>
<td>.39</td>
<td>.71</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GW-HU</td>
<td>1.00</td>
<td>.01**</td>
<td>.83</td>
<td>.21</td>
<td>.08</td>
<td>.01**</td>
<td>.84</td>
<td>.26</td>
<td>.24</td>
<td>.61</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. ANOVA Findings on the Level of Certainty and Argumentative Knowledge Construction as Reflected in Individual and Group Work (Uncertain = 1, Neutral = 2, and Certain = 3; df1 = 2, df2 = 15)
<table>
<thead>
<tr>
<th></th>
<th>Individual Ideation</th>
<th></th>
<th>Group Work</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hedging Used:</td>
<td></td>
<td>Hedging Used:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uncertain = 9,</td>
<td></td>
<td>Neutral =</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neutral = 6, and</td>
<td></td>
<td>Certain = 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Certain = 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Argument</td>
<td>.78 .44 4.25 &lt; .05*</td>
<td></td>
<td>.211 1.54 .07 &gt; .05</td>
<td></td>
</tr>
<tr>
<td>Qualifier</td>
<td>.33 .50 1.54 &gt; .05</td>
<td></td>
<td>2.11 1.36 .53 &gt; .05</td>
<td></td>
</tr>
<tr>
<td>Counterargument</td>
<td>.56 .53 .49 &gt; .05</td>
<td></td>
<td>1.11 .33 .24 &gt; .05</td>
<td></td>
</tr>
<tr>
<td>Integrated Replies</td>
<td>.11 .33 .72 &gt; .05</td>
<td></td>
<td>.00 .00 .64 &lt; .05*</td>
<td></td>
</tr>
</tbody>
</table>

References


Latifi, S., Noroozi, O., & Talaei, E. (2020). Worked example or scripting? Fostering students’ online argumentative peer feedback, essay writing and learning. *Interactive Learning Environments, 1*-15.


ALT
ChatGPT’s Performance in Spreadsheets Modeling Assessments based on Revised Bloom’s Taxonomy

Michelle CHEONG
School of Computing & Information Systems, Singapore Management University
michcheong@smu.edu.sg

Abstract: ChatGPT has taken the education scene by storm and caused uneasiness among educators. Mixed reactions were observed with some institutions banning it, while others embracing it with caution. This paper evaluates the performance of ChatGPT on solving spreadsheets modeling assessment questions with multiple test items categorized according to the revised Bloom’s taxonomy, to discover the accuracy of the answers provided at each cognitive learning level. The insights obtained may be useful for educators to design future assessment questions which focus more on testing critical thinking skills to assess the students accordingly to achieve the intended learning outcomes, and we propose recommended actions on how to do so. Our proposed methodology can be applied to other course modules to achieve their respective insights for future assessment designs and actions.

Keywords: ChatGPT, Performance Evaluation, Spreadsheets Modeling, Assessments, Revised Bloom’s Taxonomy

1. Introduction

ChatGPT (Generative Pre-trained Transformer) is a Generative AI (GAI) Large Language Model (LLM) developed by OpenAI (OpenAI, 2022) trained to provide answers across a myriad of domains. Released in November 2022 as free access to all users, New York City (The Guardian, 2023) and New South Wales and Queensland in Australia (ABC News, 2023) decided to ban it in schools with many other cities following suit. Such a reaction is reasonable as CNN reported that ChatGPT was able to pass the law exam and a business school exam (CNN Business, 2023). However, when ChatGPT was tested on three years of Singapore’s Primary School Leaving Examinations (PSLE) for 6th grader, it failed miserably in Mathematics and Science, and only managed a borderline pass for English (Business Insider, 2023). Many papers supported using ChatGPT for teaching and learning, while at the same time expressed many concerns (Alkaissi & McFarlane, 2023; Gilson et al., 2023; Kung et al., 2023). Thus, it is not clear to many whether ChatGPT, or any other GAI like Google Bard or Amazon Bedrock, is really a boon or bane to education.

In this paper, we review the recent literature on ChatGPT’s implications on education and how it performed in terms of educational assessments. We then evaluate ChatGPT’s performance in quiz questions for a university spreadsheets modeling module where the questions with multiple test items are mapped according to the revised Bloom’s taxonomy. With the answers provided by ChatGPT, we establish how well ChatGPT tackled technical questions on the different cognitive learning levels. We hope to provide educators with insights to design future assessments on spreadsheets modeling module and propose recommended actions on how to do so, especially for institutions which are ready to embrace ChatGPT in education. Our proposed methodology can be applied to other course modules to achieve their respective insights for future assessment designs and actions.
2. Literature Review

Due to the recent rapid increase in the popularity of ChatGPT, several academic papers were published on its implications on education and its performance in educational assessments. Lim et al. (2023) presented four paradoxes of GAI and their implications to the future of education. “Paradox 1: GAI is a friend yet a foe” highlighted its advantages in accelerated learning, discovery of new knowledge, reform the design of assessments to achieve higher level, and yet concerned with misinformation and its impact on academic integrity. “Paradox 2: GAI is capable yet dependent” discussed the need to make the best use of it to maximize returns, and yet be cognizant that it is dependent on the quantity and quality of prompts and its prior training. “Paradox 3: GAI is accessible yet restrictive” discussed the need for a sustainable model to promote equitable access to all. “Paradox 4: GAI is popular when banned” promoted its use in education rather than banning it which may lead to many negative consequences. Through the four paradoxes, they provided guidelines for practices and directions for future research.

Pavlik who co-authored with ChatGPT (2023) studied its implications on journalism and media. ChatGPT could provide high-quality answers and thus may pose a threat to journalists and media professionals. He suggested educators to use it to develop course but warned that students might use it in their own academic work compromising academic integrity. Alkaissi and McFarlane (2023) presented two medical cases and asked ChatGPT to provide scientific writing on specific medical and non-medical topics. The response provided was mostly factually correct but with some errors. The references provided were non-existent, demonstrating that it suffered from “artificial hallucination” where seemingly realistic outputs do not correspond to any real-world inputs. Thus, they advocated using AI output detectors in editorial processes and clear disclosures by authors. Similarly, Anderson et al. (2023) also requested ChatGPT to write two academic papers on using AI for scientific manuscripts and in sports medicine. They also found that the bibliographies generated were wrong. One concerning finding was that when using GPT-2 Output Detector to assess the originality score for both papers, the scores increased tremendously by simple paraphrasing.

In terms of ChatGPT’s performance in educational assessments, Gilson et al. (2023) and Kung et al. (2023) both evaluated ChatGPT’s performance in the United States Medical Licensing Exam (USMLE). Gilson et al. (2023) used questions from question banks for Step 1 and Step 2 exams, while Kung et al. (2023) selected sample exam questions from June 2022 to test ChatGPT’s performance on answering original questions instead of retrieving answers from the corpus which it was trained on. Both studies concluded that ChatGPT passed the threshold of 60% and can potentially assist learners in medical education. One interesting finding from Gilson et al. (2023) was that the accuracy of the responses significantly decreased as the difficulty of the question increased. They measured the types of errors committed by ChatGPT and found that it committed mostly logical errors (failure to convert information to answer), then information errors (failure to identify key information), and very few statistical errors (arithmetic error).

With ChatGPT performing reasonably well in educational assessments, how should future assessments be changed? Stokel-Walker (2022) suggested the possibility of removing essay writing altogether. Zhai (2022) suggested educators to focus on improving students’ creative and critical thinking skills instead of general skills, and to use AI tools to conduct subject-domain tasks. Echoing this, O’Connor who co-authored with ChatGPT (2023) recommended to use oral presentation and objective structured clinical examinations to increase the diversity of assessments used in nursing education.

In this paper, we evaluate the accuracy performance of ChatGPT to solve technical questions on spreadsheets modeling and each question with multiple test items was mapped against the revised Bloom’s taxonomy. Our work is different from previous works in two main areas. Firstly, past works mostly evaluated ChatGPT’s performance in non-technical modules, while we used a technical module in spreadsheets modeling. The only other work that used a technical module is the unpublished work by Malinka et al. (n.d.) where they evaluated ChatGPT’s performance in computer security modules. They only shared the types of assessments used to test ChatGPT but did not share any details of the questions.
They found great variability in the correctness of ChatGPT’s answers and artificial hallucination also occurred. Secondly, none of the previous works linked ChatGPT’s performance on test items categorized on a cognitive learning scale, while we mapped test item against the well-accepted revised Bloom’s taxonomy to achieve insights to advise future assessment designs in response to recommendations by Zhai (2022) and Lim et al. (2023) to focus more on assessing critical thinking skills.

3. Bloom’s Taxonomy and Spreadsheets Modeling Module

3.1 Bloom’s Taxonomy

Bloom’s taxonomy is a theoretical model (Bloom, 1956) intended to classify cognitive learning from simple to complex level in a progressive manner, to aid educators in reducing duplicate assessment items to achieve the same learning outcome. There are six levels (1) knowledge, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6) evaluation. For each level, there are action verbs to classify assessments accordingly. To represent the 21st century teaching and learning, Anderson and Krathwohl (2001) revised Bloom’s taxonomy into a two-dimensional model. The six levels remained but they were changed from noun to verb as (1) remember, (2) understand, (3) apply, (4) analyze, (5) evaluate, and (6) create. Levels 5 and 6 were swapped, and a second dimension called the “knowledge dimension” which has four levels (1) factual, (2) conceptual, (3) procedural, and (4) metacognitive, was added in addition to the original “cognitive dimension”.

One main challenge in applying Bloom’s taxonomy stems from the overlapping action verbs in multiple levels, leading to ambiguity in its application as highlighted by Das, Mandal, and Basu (2022). Even with the revised taxonomy, the ambiguity issue did not go away totally. In this paper, we classify our test items to the different levels of the revised Bloom’s taxonomy based on two module instructors’ expertise in the subject matter, and a third instructor was consulted in case of different classification.

3.2 Spreadsheets Modeling Module

We use two quizzes created for a university level module in spreadsheets modeling to evaluate ChatGPT’s performance. The module objective is to teach students how to build spreadsheet models from scratch and perform calculations and data analysis to support business decision making. The learning outcomes include formulating business problems and integrating business analysis skills to model and appraise business problems; acquiring computer skills to perform problem analysis; and acquiring competency in spreadsheets tool. The module covers six main topics as described in Table 1.

Table 1. Six Main Topics Covered in the Spreadsheets Modeling Module

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Basic modeling techniques to formulate arithmetic equations and using Excel functions to perform basic statistical calculations.</td>
</tr>
<tr>
<td>2</td>
<td>Spreadsheet engineering skills in terms of using conditions to evaluate outcomes, create basic charts, compute intercept and slope of linear lines, create different order of polynomial functions to represent business logic relationships to be included into the model to seek for best answer using Goal Seek.</td>
</tr>
<tr>
<td>3</td>
<td>Financial calculations using Excel functions to understand and compute time value of money to assess investment options and returns.</td>
</tr>
<tr>
<td>4</td>
<td>Data lookup and linkup to handle changing inputs which affect outputs, and to determine optimal solutions to business problems which are subjected to constraints using Solver Add-in.</td>
</tr>
</tbody>
</table>
3.3 Assessment Quizzes

The module instructor designed the two quizzes for the spring term of academic year 2022/23. As the quiz questions were original, they did not exist before January 2023, and thus will not form part of the ChatGPT’s training corpus. Each quiz describes a business scenario and contains multiple questions as listed in Table 2. Q1 to Q5 are for Quiz 1, while Q6 to Q10 for Quiz 2. Each question contains multiple test items, and each item is mapped to the revised Bloom’s taxonomy (BT), with the lowest at 3, and the highest at 6. It is not common to test students on levels 1 and 2 for university level modules. The mapping was done by two instructors and the inter-rater agreement between them was evaluated using the Cohen’s kappa statistics (Hallgren, 2012), and found to be 0.902.

Table 2. Questions and Test Items from Two Quizzes and their Bloom’s Taxonomy (BT)

<table>
<thead>
<tr>
<th>Question</th>
<th>Test item</th>
<th>BT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q1. A mall offers different discounts, Types A to F, and allows shoppers to split their total purchase amount into multiple payments to enjoy as many discounts as possible. Type A is a $10 discount with minimum $140 spending. Type B is a $20 discount with a minimum $500 spending. Type C is a $10 discount with a minimum of $125 spending. Type D is a $45 discount with a minimum spending of $800. Type E is a $15 discount with a minimum $800 spending. Type F is a 1.5% discount with no minimum spending. Compute the ratio defined as discount divided by minimum spending, and compute the rank of the ratio in descending order using a suitable Excel function. Create an Excel spreadsheet model, called Table 1, to determine the answer.</strong></td>
<td>• Test arithmetic to compute ratio.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>• Test identification and application of correct Excel function to rank ratio.</td>
<td>4</td>
</tr>
<tr>
<td><strong>Q2. You plan to buy a computer for $1200 and wish to split the purchase amount into multiple payments to use as many discounts as possible, to enjoy the maximum discount. Create a second Excel spreadsheet model, called Table 2. In the first column, list the rank from 1 to 6. Then use any Excel lookup function to retrieve the discount types from Table 1 corresponding to the rank position, displaying the discount, the amount to be charged to this discount, and the discount enjoyed. It is only smart to charge the minimum amount to</strong></td>
<td>• Test identification and application of lookup function to retrieve data.</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>• Test formulation of complex formulas to determine if discount is used and amount charged to each discount considering all</td>
<td>6</td>
</tr>
<tr>
<td>Question</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Q3.</td>
<td>With the final total amount computed earlier, you can choose to pay it over three equal monthly instalments. Using an annual interest rate of 2%, what would be the present value of the final total amount?</td>
<td></td>
</tr>
<tr>
<td>Q4.</td>
<td>Your friend wants to buy your computer for $1400 today. How much would you earn today if you were to sell your computer to your friend, based on the present value of the final total amount you paid computed earlier?</td>
<td></td>
</tr>
<tr>
<td>Q5.</td>
<td>At what annual interest rate would you be able to earn $300 today if you were to sell your computer to your friend for $1400, based on the present value of the final total amount you paid using this new annual interest rate?</td>
<td></td>
</tr>
</tbody>
</table>
| Q6.      | At a clinic, patients join a single queue lining up one after another to see the doctor. Assuming the inter-arrival time of patients follows an exponential distribution with mean 4 minutes. Out of these patients, 35% of them see the doctor due to suspected COVID-19 infection, while the remaining 65% are normal patients. For those who suspect COVID-19 infection, the doctor will test to confirm if they are indeed infected. It was found that 70% of them are confirmed. Due to different medical conditions, the consultation time differs, and they all follow normal distributions with different means and standard deviations. For the normal patients, the mean is 5 minutes and standard deviation is 1 minute. For the suspected and not confirmed patients, the mean is 7 minutes and standard deviation is 2 minutes. For the suspected and confirmed patients, the mean is 15 minutes and standard deviation is 3 minutes. Create an Excel model Table 1 with the first column to simulate the

---

- Test amount payable if discount is used.
- Test computation of total amount payable.
- Test identification and application of correct Excel function to perform financial calculation.
- Test arithmetic to compute earnings.
- Test recursive calculation using Goal Seek or Solver.
- Test data simulation using exponential distribution for inter-arrival time.
- Test conditional probability to simulate patient type.
- Test time calculation.
- Test IF condition to get correct inputs for simulating consultation time using normal distribution for different patient type.
inter-arrival time of 50 patients, then compute the arrival time of the patients using 9am as the reference starting time in the second column. Then add a new column to simulate if this patient is suspected COVID-19 case using “Yes” or “No” label. If “Yes”, then add another column to simulate if this patient is a confirmed COVID-19 case using “Yes” or “No” label. Finally, add another column to simulate the consultation time depending on the patient type.

<table>
<thead>
<tr>
<th>Q7.</th>
<th>Continuing with Table 1 created earlier, for each patient determine the consultation start time, consultation end time, the wait time, and the system time. Update as Table 2.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Test queue system concepts.</td>
</tr>
<tr>
<td></td>
<td>• Test arithmetic to compute consultation time, end time, wait time system time.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q8.</th>
<th>While waiting in the line, it is possible that a non-COVID-19 patient (e.g. Person B) may get infected by the person standing directly in front (e.g. Person A), and/or the person standing directly behind (e.g. Person C), if Person A and/or Person C are confirmed COVID-19 cases, and the overlap time they spend standing in line next to Person B exceeds 30 mins. Continuing with the Table 2 created earlier, add two new columns to determine the overlap time between Person A and Person B, and overlap time between Person B and Person C. Finally, add a last column to indicate for each patient, will this patient get infected using “Yes” or “No” label. However, if this patient is already a confirmed COVID-19 patient, indicate “C+” instead.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Test queue system concepts.</td>
</tr>
<tr>
<td></td>
<td>• Test arithmetic to compute overlap time.</td>
</tr>
<tr>
<td></td>
<td>• Test formulation of complex formulas to determine new infection status taking into account multiple linked considerations.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q9.</th>
<th>Using the updated Table 2 created earlier, compute the number of possible new infections, and use Data Table to repeat this simulation 500 times to determine an overall average number of possible new infections.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Test identification and application of correct Excel function to count based on test condition.</td>
</tr>
<tr>
<td></td>
<td>• Test Data Table concept to repeat simulations.</td>
</tr>
<tr>
<td></td>
<td>• Test arithmetic to compute average.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q10.</th>
<th>The doctor thinks that he should reduce the mean consultation time by 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Test Data Table concept to repeat</td>
</tr>
</tbody>
</table>
He can choose to do that for only one of the cases, normal patients, suspected and not confirmed patients, or suspected and confirmed patients. Regardless of which case he decides to reduce the mean consultation time, the wait time for all patients will be reduced, and thus lead to reducing the number of possible new infections. Determine the new overall average number of possible new infections for each case using Data Table and conclude which case will be most effective and explain why.

To simulate the similar answering process as real students, the questions were posted to ChatGPT sequentially in one single continuous conversation to allow it to refer to its previous responses to generate the next response for follow-up questions. This is an in-built capability that ChatGPT has, which fascinated many users.

4. ChatGPT’s Performance in Spreadsheets Modeling Assessment

We evaluate ChatGPT’s performance for the quizzes over three complete runs and compute the average accuracy at each cognitive level. While the responses from ChatGPT were not exactly the same for each run, they were rather similar albeit different mistakes committed. The main difference comes from the way the responses were presented, which could be in the form of step-by-step explanation or presented as table. We found that ChatGPT can perform reasonably well at level 3 with 67% accuracy and the accuracy decreases as the cognitive level increases. This result is similar to the work by Gilson et al. (2023) where they reported a significant decrease in accuracy as the question difficulty increases for USMLE.

Table 3. ChatGPT’s Accuracy Performance at Different Bloom’s Taxonomy (BT) Level

<table>
<thead>
<tr>
<th>BT</th>
<th># items</th>
<th>ChatGPT Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 (apply)</td>
<td>6</td>
<td>67%</td>
</tr>
<tr>
<td>4 (analyze)</td>
<td>9</td>
<td>39%</td>
</tr>
<tr>
<td>5 (evaluate)</td>
<td>5</td>
<td>33%</td>
</tr>
<tr>
<td>6 (create)</td>
<td>3</td>
<td>22%</td>
</tr>
</tbody>
</table>

At level 3 (apply), ChatGPT was able to perform arithmetic calculations reasonably well. However, it committed mistakes when it misinterpreted the information provided in the question. For example, in Q1, the 1.5% discount for Type F is equivalent to a discount ratio of 0.015 but ChatGPT misinterpreted it and indicated it as N/A instead (Figure 1).

![Figure 1. ChatGPT Misinterprets Information.](image)
At level 4 (analyze), ChatGPT was able to analyze the problem and determine the correct Excel function to use but was not able to apply the function correctly. For example, in Q3, it correctly identified using PV function, but the arguments to fill in the function were wrong, where the \([\text{type}]\) should be set as 1 and not ignored which means type 0 (Figure 2).

\[
=\text{PV}(\text{rate}/12, nper, \text{pmt}, [fv], [\text{type}])
\]

To display the present value answer in Table 2, we can add a new row below the Amount Payable row, and in the Amount Payable column for that row, enter the following formula:

\[=\text{PV}(0.002, 3, -H6/3)\]

*Figure 2. ChatGPT Correctly Identifies PV Function but Filled in Wrong Argument.*

At level 5 (evaluate), ChatGPT was able to explain the evaluation approach, but was not able to perform the evaluation correctly. For example, in Q9, ChatGPT can explain how to use Data Table to perform the simulation, but the actual calculation using COUNTIFS function was not done correctly, as it should only count patients with new “Yes” label, and not those who are “No” and “C+” labels. (Figure 3).

\[=\text{COUNTIFS}(\text{Table 2!S}2:\text{S}51, "No", \text{Table 2!S}2:\text{S}51, "C+", \text{Table 2!S}2:\text{S}51, "Yes")\]

*Figure 3. ChatGPT can Explain Data Table Setup but Performed Wrong Calculation.*

At level 6 (create), where complex formulas which involve linking several parts of calculations together are expected, ChatGPT failed miserably. It tends to over-simplify the calculations leading to the wrong answers. For example, Q2 required formulas to determine if a specific discount is used and the amount charged to each discount if that discount is used. This requires all discounts to be considered together as a whole to maximize total discount, and not treat each discount as separate consideration.

### 5. Insights and Implications
We found that ChatGPT’s responses are often self-contradictory and incoherent when responding to technical calculation questions. It can describe the calculation process correctly, but the actual calculation was performed incorrectly. Very often, the suggested calculation steps are messy and cell references in the provided formulas are not referring to the correct cells. One plausible explanation for such a behavior is that ChatGPT is a language model. It excels in “auto-complete” and thus performed well in essay writing. However, when it comes to logical and coherent thinking required in complex calculations involving linking multiple considerations, it was not able to handle.

On the other hand, we found that ChatGPT is knowledgeable in explaining Excel functions in terms of the arguments and how to use them. And it is surprisingly good in queuing system, probability and distribution functions concepts. It could be due to the corpus it was trained on which is mainly based on textual content and facts.

Generative AI is here to stay and will become more powerful creating immense impact on education. Instead of outright banning it, educators should consider how to best use it for course development, teaching and assessment, as echoed by many researchers including Lim et al. (2023) and Zhai (2022). Based on the results of our study, we provide three recommendations for educators who teach spreadsheets modeling modules. One, allow students to use ChatGPT to answer in-class activity questions (not actual assessments), and guide them to identify errors and limitations in the answers and to suggest how to improve the answers. For example, request students to build a first model based on ChatGPT’s suggested answers, and then build a second model with any identified errors corrected. Two, ask students to formulate questions to ask ChatGPT and through analyzing the answers, learn to sharpen the questions to represent the exact intent of the questions as a form of interactive learning (Rospigliosi, 2023). For example, students can ask ChatGPT to compare two different investments options, and then sharpen the question to ask for comparison using net present value or internal rate of return calculations. And lastly, design assessment questions which focus on testing higher order thinking skills, with test items from level 4 onwards (Zhai, 2022). For example, design questions that link data lookup function with conditional test to determine the answer, instead of questions that only require simple arithmetic calculations like computing the sum or average. Our recommendations aim to overcome diminishing learners’ own innovative capacities and critical thinking skills, and focus on enhancing their higher order thinking skills.

6. Conclusions and Future Research

Our study shows that ChatGPT’s performance in spreadsheets modeling assessment questions was good up to level 3 of the revised Bloom’s taxonomy, and its accuracy decreases as the cognitive level increases. It could explain technical facts very well and performed reasonably well in basic arithmetic calculations. However, from level 4 onwards, ChatGPT was unable to apply Excel functions correctly even though it has identified the correct Excel function; unable to evaluate complex situations even though it could explain the evaluation approach; and unable to formulate complex formulas that link multiple considerations as its formulations were often simplified. However, we believe that ChatGPT and other GAI will become better and as educators we should exploit their full potential in education while keeping in mind how to overcome the concerns on cheating, plagiarism and loss of independent thinking skills in students. This calls for a balanced approach and a paradigm shift in education.

We recommend several future research possibilities involving GAI. From the instructors’ perspective, we can perform research in using GAI to generate assessments and assess the cognitive levels of such assessment questions to understand if GAI can generate questions of higher order thinking skills, and how much GAI can lighten instructors’ workload. From students’ perspective, we can quantify the improvement in knowledge gain in students when using GAI as part and parcel of their learning process. From the GAI tool’s perspective, we can consider developing and implementing customized GAI tool using LLM
to teach and learn technical modules such as spreadsheets modeling or Python programming, and assess the tool’s effectiveness.

References


Promoting Middle School Students' Achievement and Attitude toward Science Learning through Sphere Recognition-Based AR Application

Ruixue LIUa*, Lijun LIANGa & Xiaodong WEIa

aSchool of Educational Technology, Northwest Normal University, China

*isnow0211@163.com

Abstract: With the advancement of Augmented Reality (AR) technology, considerable studies have explored AR’s application in various educational fields, especially in the field of science. This paper used a sphere recognition-based identifier to develop an AR application that supports middle school students learning of science. A quasi-experimental method was adopted, using a pretest and posttest of science knowledge to measure academic performance. A questionnaire was conducted to measure students’ attitudes toward the use of AR in science lessons. The research sample included 80 middle school students and was divided into experimental and control groups. Students in the experimental group learned science lessons by using a sphere recognition-based AR application, while the control group learned the same material by using a traditional method. Results showed that students who learned science lessons in a sphere recognition-based AR application performed better in academic achievement than that in the control group. In addition, the experimental group showed a positive attitude toward the use of AR in science classes. This study provides empirical evidence to explore the use of sphere recognition-based AR in science education.

Keywords: Augmented reality, academic achievement, middle school, science learning, attitude

1. Introduction

Science learning is an important subject in the educational area. However, learners often face difficulty in learning science because it involves many complex and abstract concepts (e.g. gravity, electromagnetic phenomena, etc) (Anderson & Barnett, 2013). In this way, learners may have misunderstandings in the process of understanding and analyzing scientific phenomena due to the lack of intuitive and vivid presentation. In addition, real scientific practice faces problems such as geographical distance and security risks (Zhang & Wang, 2021). For example, previous studies have indicated that students could grasp the description of the phases of the moon, but they often misunderstand the interpretation of the moon phases and believe that the moon phases are formed by the occlusion of the sun (Lelliott & Rollnick, 2010; Wilgenbus & Lena, 2009). Therefore, many students may have difficulties in science learning, such as low interest and confidence in science curricula, and low learning achievement.

With the development of emerging technologies, various visual technologies such as virtual reality, and augmented reality can be applied in educational fields to promote student’s learning experience. Augmented reality (AR) is a technology that seamlessly connects virtual information with the real world and has the characteristics of virtual and real fusion (Azuma, 1997). It provides opportunities for learners to observe and explore 3D models and support their understanding of intellectual concepts (Xu et al., 2022). With the help of this technology, learners can place 2D or 3D virtual information in real scenes by using their devices to identify markers. Moreover, AR has the advantages of presenting...
abstract scientific knowledge and supporting scientific exploration. Fleck and Simon (2013) indicated that AR is a good scaffold for learning science knowledge due to its ability to enhance sensorimotor interactions, visual guides, and realistic representation. This study aims to investigate the effect of sphere recognition-based AR application on middle school students learning of science. The research questions are as follows:

- Compare to the traditional teaching methods, do students who learned science through using a sphere recognition-based AR application perform better in academic achievement?
- What are the attitudes of students toward the use of sphere recognition-based AR applications in science learning?

2. Literature Review

2.1 Science Learning

Science is a discipline concerned with the study of organisms and events in nature. It requires people to form concepts and principles by observing and generalizing things and phenomena (Karagozlu, 2021). The purpose of science education is that it requires students to form scientific concepts, master the nature of science, and improve their willingness to actively participate in scientific learning (National Research Council, 1996). One of the main challenges for science education is helping learners to develop a conceptual understanding of scientific concepts (Osborne et al., 2016). However, there is a lack of physical conditions to support learners in practicing learning tasks in a school classroom setting. To solve these problems, researchers have attempted to integrate AR and virtual reality into the science curriculum (Monita & Ikhsan, 2020). Since these technologies provide opportunities for the visualization, simulation of scientific knowledge, and science practice.

2.2 AR in Science Learning

AR is a technology that can superimpose virtual information in the real world. It has three characteristics of the combination of reality and reality, real-time interaction, and 3D registration (Azuma, 1997). Based on AR’s characteristics, it brings new opportunities in the field of education (Zhang et al., 2022). First, the combination of reality and reality can present 3D learning content for learners, which can contribute to the visualization of knowledge and solve the problems of simulating inaccessible, invisible, and dangerous scientific phenomena. Second, AR supports students’ real-time interaction with virtual information in a real environment. Moreover, some studies have shown that AR plays an important role in increasing learning motivation and engagement, as well as enhancing learners’ creativity and imagination (Demircioglu et al., 2022; Lorusso et al., 2018).

Previous studies have applied AR in science education. For example, Liu et al. (2023) developed an AR application for chemistry courses for middle school students level. The results showed that AR helped increase knowledge gains and improved motivation. Wang (2022) developed a game-based AR science learning system for middle school students to learn electromagnetic phenomena. Results found that the proposed system has advantages in improving interactivity, learning performance, and learning satisfaction. Tina et al. (2019) developed a real-world-oriented smartphone AR system and they found that the system is an effective learning tool and improves the interest in learning science.

Overall, previous studies have demonstrated the advantages of AR in science education. Among them, marker-based (such as image-based) AR is commonly used. Other types of AR, such as location-based and object recognition-based AR, are less implemented. In this study, we use an AR application based on sphere recognition to explore its effects on science learning.
3. Development of a Sphere Recognition-based AR Science Learning System

3.1 System Structure Overview

In this study, a sphere recognition-based AR application was developed using Unity and Vuforia engines. The process of the development of the AR application requires the use of “Scanner” software to scan spheres and then generate data files of the sphere model. Following that, upload the data file to the database of the Vuforia engine. Next, connect the corresponding Vuforia engine database in the Unity engine, and edit the sphere data in the Unity engine to form the final AR application. The AR application was eventually released as an APK and installed on students' mobile devices. Figure 1 shows the system structure.

![Figure 1. The System Structure of the AR Application.](image)

3.2 Sphere Recognition-based AR Science Learning Application

The theme of the science curriculum is “Exploring the Earth-Moon System”. The AR application consists of three main parts: (a) “Moon Observation”, (b) “Exploring the Moon Phase”, and (c) “The Solar and Lunar Eclipses” module. Students can click the module selection button to enter the interface according to the learning progress on the AR application homepage (see Figure 2).

![Figure 2. The Interface of the Homepage](image)
The “Moon Observation” module includes the moon’s position in the universe, the
types of terrain on the moon, and the characteristics, and locations of the terrain (see Figure
3). Students scan the sphere using their mobile phones, and a 3D model of the moon
appears on the page. Students can interact with the 3D model, including zooming and
rotating. The surface of the 3D model is equipped with marker points, and students are
guided by the marking points. When students click on the markers, text and images
describing the model appear on the right side of the screen. In addition, by clicking on the
buttons at the bottom of the screen, students can get an audio or video explanation of the
lunar terrain.

Figure 3. The Interface of the “Moon Observation” Module

The “Exploring the Moon Phases” module includes the types of moon phases, the
evolution of moon phases, and reasons for the formation of moon phases (see Figure 4).
Students scan the sphere using their mobile phones, and a 3D model of the Earth-Moon
system appears on the page. After the student selects the date on the left side of the page,
the 3D model shows the relative positions of the Earth and the Moon, and an image of the
corresponding moon phase appears on the right side of the page. Students can click the
buttons at the bottom of the screen to get text, audio, and video explanations of the moon
phase.

Figure 4. The Interface of the “Exploring the Moon Phases” Module
In the “The Solar and Lunar Eclipses” module (see Figure 5), students learn the types and patterns of solar and lunar eclipses and analyze the reasons for their formation. They select an option from the drop-down menu in the upper right corner of the screen. The 3D model is corresponding to the options that will be presented on the page. Students select the “Annular Solar Eclipse” option, and the 3D model on the page simulates the relative positions of the Sun, Moon, and Earth during an annular solar eclipse, as well as the light and shadow phenomena. Text and images of the annular solar eclipse are presented at the bottom of the page. Students can also use the buttons on the left side of the page to access audio and video explanations of the annular solar eclipse.

Figure 5. The Interface of the “Solar and Lunar Eclipses” Module

4. Methods

4.1 Research Design

This study used a quasi-experimental method to investigate the effect of a sphere recognition-based AR application on students’ academic performance and attitude toward science learning. The two classes participating in the experiment were divided into experimental and control groups. Before the experiment, both groups of students took a pretest to ensure that they had a similar knowledge level. Following that, students in the experimental group learned science lessons by using a sphere recognition-based AR application, while the control group learned the same content using traditional teaching methods. After finishing the experiment, a posttest and questionnaires were administered to both students.

4.2 Participants

The study was conducted in a public middle school in the Northwest region of China. A total of 80 students from two classes with an average age of 13 years old participated in this study. Among them, 40 students were in the experimental group and 40 students were in the control group. Both classes had approximately the same proportion of boys and girls. The science lessons were taught by the same teacher with more than two years of teaching experience. All students had not been exposed to AR science lessons before the experiment.
4.3 Measurement Tools

4.3.1 Pretest and Posttest

To test students’ science learning effect, pretest, and posttest were used to measure students’ academic achievement before and after the learning activity. Both tests were developed by research and the teacher and have a Cronbach alpha of 0.83. The pretest and posttest consist of 10 multiple-choice items (e.g., When does the lunar eclipse occur? A. new moon B. upper moon C. lower moon D. full moon), 6 true or false items (e.g., The average distance from the moon to the earth is about $3.84\times10^8$ km. A. True B. False), and 2 short-answer items (e.g., How many stages are there in the moon phase? What is the formation principle of the moon phase?). The perfect score on both tests is 100. An expert panel of three educators examined the validity of the tests. Both science teachers and the expert panel reviewed the final version of the tests. We modified the inappropriate items in the tests to ensure high reliability and validity.

4.3.2 Questionnaire

This study used an AR Application Attitude Scale (ARAAS) developed by Kucuk et al. (2014) to measure students’ attitudes toward the use of AR applications in science lessons. The ARAAS adopts a 5-point Likert scale (‘1’= Strongly Disagree, ‘5’= Strongly Agree) and consists of 14 items with three dimensions, including “using satisfaction”, “using anxiety”, and “using willingness”, among which six positive descriptions reveal students’ satisfaction with using AR application, (e.g., I enjoy the AR lessons ), six negative descriptions reflect students’ anxiety level in using AR applications (e.g., It is difficult to use AR applications), two positive descriptions show students’ intention to use AR applications in the future (e.g., I want AR applications in other lessons). The values of Cronbach’s alpha for the three dimensions were 0.86, 0.82, and 0.89, respectively. The value of Cronbach’s alpha for the overall scale was 0.83, indicating that the scale is reliable.

4.4 Procedure

As shown in Figure 6, the experiment lasted for 5 weeks. In the first week, students in both groups took a pretest. From the second to the fourth week, the students participated in science lessons. Each week, both groups of students took a 45-minute science lesson. An example of a 45-minute class is shown in Table 1. During the lesson, students in the experimental group learned science lessons using a sphere recognition-based AR application. the teacher first conducts the lesson by introducing the topic. After that, the teacher introduces the use of AR and guides the students to use it. When finished, students use the AR application for science learning. In the end, the teacher guides the students to summarize and share the knowledge learned in the lesson. The control group followed the same teaching process and used traditional teaching methods. In the fifth week, both groups took the posttest. Meanwhile, students in the experimental group were also required to complete the questionnaire.
4.5 Data Analysis

In this study, data were analyzed using SPSS software. Descriptive analyses and t-tests were used to analyze the student’s academic achievement. In terms of attitudes toward AR technology use, descriptive analysis was used to analyze data from students’ ARAAS questionnaires.

5. Results

5.1 Academic Achievement

The results of the pretest scores are given in Table 2. The Mean values of the pretest scores were 47.75 for the experimental group and 47.50 for the control group respectively. The t-test results ($t = 0.27$, $p > 0.05$) showed no significant difference between the two groups, indicating that both groups had a similar level of a priori knowledge before the learning activity.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td></td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.79</td>
</tr>
<tr>
<td>Experimental</td>
<td>40</td>
<td>47.75</td>
<td>4.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>40</td>
<td>47.50</td>
<td>4.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posttest</td>
<td></td>
<td></td>
<td></td>
<td>9.34</td>
<td>0.00*</td>
</tr>
<tr>
<td>Experimental</td>
<td>40</td>
<td>83.70</td>
<td>4.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>40</td>
<td>72.33</td>
<td>6.25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.01.
At the end of the learning activities, both groups of students took a posttest. The Mean values of post-test scores were 83.70 for the experimental group and 72.33 for the control group. The \( t \)-test results \( (t = 9.34, p < 0.01) \) showed a significant difference between the two groups. It can be concluded that the students who use AR applications for science learning had better learning performance.

### 5.2 Attitude Towards AR Applications

As shown in Table 3, the Mean values and SD of the overall attitudes for the experimental group were 3.95 and 0.27, indicating that students in the experimental group gave high attitudes toward the use of AR applications in science classes. Moreover, the Mean values and SD of satisfaction for the experimental group were 3.91 and 0.38, and the willingness for the experimental group was 3.86 and 0.57. In addition, the Mean values and SD of anxiety for the experimental group were 1.99 and 0.29. The results showed that students were satisfied with using AR applications in their courses, with a low level of anxiety and they generally showed their intention to continue using AR applications in the course. It can be concluded that students in the experimental group have positive attitudes toward using AR applications in science courses.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td>40</td>
<td>3.17</td>
<td>5.00</td>
<td>3.91</td>
<td>0.38</td>
</tr>
<tr>
<td>Anxiety</td>
<td>40</td>
<td>1.00</td>
<td>2.50</td>
<td>1.99</td>
<td>0.29</td>
</tr>
<tr>
<td>Willingness</td>
<td>40</td>
<td>3.00</td>
<td>5.00</td>
<td>3.86</td>
<td>0.57</td>
</tr>
<tr>
<td>General ARAAS</td>
<td>40</td>
<td>3.50</td>
<td>5.00</td>
<td>3.95</td>
<td>0.27</td>
</tr>
</tbody>
</table>

### 6. Discussion and Conclusion

In this study, we explored the effects of a sphere recognition-based AR application on middle school students’ academic performance and their attitudes toward science learning. The results showed that students who used a sphere recognition-based AR application for science learning had significantly higher academic achievement than that of the control group. In addition, students in the experimental group gave positive attitudes regarding the satisfaction and willingness to use AR applications in a science curriculum.

In terms of academic achievement, the results of the current study are consistent with previous studies (Kalemkus & Kalemkus, 2022; Wang, 2022). In this study, the reasons for such an effect may be as follows. First, AR helps visualize scientific knowledge. Although middle school students have certain abilities of cognitive and abstract, they are prone to misunderstand complex scientific phenomena such as the moon phase. AR presents complex and abstract scientific phenomena in the form of 3D models in the real world, and students can observe intuitively through real-time interaction to facilitate understanding and learning. Second, this study uses spheres as identified objects because the external shapes of the Earth, Moon, and Sun are spherical. Compared with 2D images, the 3D model has a better sensory effect in the way of sphere recognition. The fun and manipulability of the learning process motivate students to maintain a high level of motivation and interest in learning, which provides the possibility for learning more knowledge and finally effectively improves their academic performance.

Moreover, students showed positive attitudes toward the use of AR applications, which confirmed other studies’ similar results (Celik & Ersanli, 2022; Fidan & Tuncel, 2019; Kaya & Bicen, 2019; Karagozlu et al., 2019). They were satisfied with the use of AR application in science lessons. Visual observation displayed by the AR helps students understand and learn from the knowledge. They also looked forward to using AR in future science lessons.
7. Limitations and Future Research

Although the above-mentioned study provides a positive result on using AR in middle school science learning, some limitations still exist in the current study. First, students were first exposed to AR learning activities in this study. Therefore, the novelty of AR learning activities may be a reason for the positive effects. In future research, pre-experiments should be performed on this issue to eliminate the influence of this factor on experimental results. Second, the experimental sample was limited to middle school students. Future studies should consider taking other age levels into account and increasing the sample size to obtain more comprehensive and accurate data. Moreover, this study used a sphere recognition-based AR application to investigate the students’ learning effect in science lessons. In future research, the effect of using AR applications of other recognition methods can be examined, and comparative research can be conducted to see which recognition method has a better effect in supporting students’ learning. It is also necessary to consider the long-term effects of the use of sphere recognition-based AR applications in middle school science learning.

Acknowledgments

The research was supported by the National Natural Science Foundation of China [grant number 62067009], and the 2023 Gansu Province University Youth Doctoral Support Project (Project No. 2023QB-110).

References


Experiential Application on Student Knowledge Gains, Learning Motivation, and Technology

Lorusso, M. L., Giorgetti, M., Travellini, S., Greci, L., Zangiacomi, A., Mondellini, M., Sacco, M., &
Renì, G. (2018). Giok the alien: an AR-based integrated system for the empowerment of


development and validation of a learning progression for argumentation in science. *Journal of
research in science teaching, 53*(6), 821-846.

smartphone AR-supported learning system based on planetarium contents for seasonal
constellation observation. *Applied Sciences, 9*(17), 3508.

Wang, Y. (2022). Effects of augmented reality game-based learning on students’

Boksenberg (Eds), *Proceedings of the International Astronomical Union* (pp. 629-641).
Cambridge University Press.

augmented reality on science learning: A meta-analysis. *Journal of Science Education and
Technology, 31*(5), 621-637.

systematic review and meta-analysis of the literature from 2000 to 2020. *Sustainability, 14*(15),
9725.

Preparation for Future Lockdowns: A Comparison of Student LMS Activity During and After COVID-19

Ezekiel Adriel LAGMAYa & Maria Mercedes RODRIGOa
aAteneo Laboratory for the Learning Sciences, Ateneo de Manila University, The Philippines
*ezekiel.lagmay@obf.ateneo.edu

Abstract: We use CausalImpact analysis to compare student online activity during the COVID-19 pandemic, when all classes were online, and post-pandemic, when classes were offered in a mix of modes. Student activity was operationalized as activity logs from the Canvas Learning Management System (LMS). When compared with student activity in pandemic fully online classes, we found that, contrary to our hypotheses, student activity in post-pandemic fully online classes decreased while activity in post-pandemic fully onsite classes increased. We attribute the decrease in participation in fully online classes to factors such as Zoom fatigue and feelings of loneliness and prolonged isolation. The increase in student online activity in fully onsite classes might be explained by students’ general preference for face-to-face interactions. They may have felt more connected to teachers and peers, leading to greater productivity overall. The discourse about the transition from online learning brought on by the pandemic is dominated by survey research. By examining student online activity in an LMS, this paper contributes an analysis of empirical data that supports the findings of prior studies about student perceptions of the online learning experience and their adjustments to the next normal.

Keywords: Learning Management Systems, COVID-19 Emergency Remote Teaching, CausalImpact

1. Introduction

In his keynote address during the International Conference for Computers in Education, Baker (2022) stated that, despite advances in analytics and data mining, the field still lacks theories “that could help build models robust to complexity and change. We still lack understanding of how change impacts models.” We have learned that models perform less well when the conditions under which they are applied change (see “Algorithmic Bias in Education [Wiki],” n.d.). There are many possible causes of model degradation including changes in user interface design, differences in the way errors are counted, and differences in the learning population over the years (Baker, 2022).

When the COVID-19 pandemic forced schools to close in early 2020, many of the prediction models and detectors stopped being useful because some of the attributes that fed into them, e.g. student attendance data, were no longer being recorded or were recorded in different ways. As we emerge from the pandemic, educational systems have retained some of these new policies and formats, such as a greater openness to hybrid formats. In this next normal, can we still trust models that were built on pre-pandemic data? What value, if any, can we derive from data collected during the pandemic?

We argue that pandemic data still offers us value because it helps us anticipate the effects of lockdowns that may take place in the years to come. Climate change, new pandemics, civil unrest, and other factors may prompt educational systems to migrate online temporarily and the return to face-to-face formats later on. In this paper, we examine learning management system (LMS) activity from two time periods: during the pandemic, when all classes were online, and when most restrictions were lifted and classes were
offered in a variety of formats. We use CausalImpact analysis (Brodersen et al., 2015) to determine differences in student LMS participation when all classes were fully online versus when they were offered in fully onsite, fully online, and variants of the two modes. We hypothesize that fully online student online participation will remain unchanged while student LMS participation for onsite classes and mixed-mode classes will decline.

2. Related Literature

As the pandemic waned, many studies emerged trying to assess the differences in learner perceptions of higher education teaching strategies, student course satisfaction, student intention to continue using e-learning platforms, and factors affecting student engagement. A study conducted at the University of Rajasthan in India (Sharma & Alvi, 2021) attempted to determine differences in student perceptions of various teaching methods before and during the pandemic. Using a questionnaire, they found that students had more positive opinions about pre-pandemic, blended e-learning formats than the purely web-based learning that they experienced during lockdown.

These findings were corroborated by the US-based study of Corral and Fronza (2022). They found that student course satisfaction declined when courses moved online. Student satisfaction ratings showed that satisfaction was high at the start of the pandemic. As the pandemic extended, though, both student satisfaction and engagement waned. They returned to normal levels when face-to-face classes resumed.

What accounts for the differences in satisfaction and engagement? Individual differences were key factors in overall satisfaction. Students whose learning preferences were compatible with distance learning and with sufficient technical resources and support services were more likely to have a more satisfying experience (Clary et al., 2022). Not all students enjoyed this level of compatibility. Many expressed that online learning was lonely and fostered feelings of laziness (Singh et al., 2021). Traditional learning, they said, made them feel more productive. Many had to cope with considerable “lifeload”—the sum of all pressures a student has to confront (Hews et al., 2022). Those experiencing external stresses engaged less than those who experienced greater well-being.

Moving forward, researchers found several types of student perceptions that influence students’ intention to engage with distance learning again (Zacharis & Nikolopoulou, 2022). These included student perceptions of e-learning as enjoyable, that their time and effort lead to worthwhile outcomes, the adequacy of technical and institutional support, and the beliefs that important others such as peers and university mentors have about the students themselves.

The consensus from prior, recent research is that students are, in general, more satisfied with face-to-face formats than online formats. While this finding may be consistent with our own intuition and experience, all of these studies tended to rely on survey data. They queried students regarding their perceptions, but they could not offer empirical data about student online participation. This paper contributes to the discourse by analyzing student participation as represented by LMS activity.

3. Methodology

3.1 Dataset Preprocessing

This study makes use of Canvas LMS log files collected from a university in Quezon City, Metro Manila, Philippines during the Summer period of Academic Years (AY) 2021-2022 and 2022-2023, which serve as the pre-intervention period and the post-intervention period respectively. This is because the classes during Summer of AY 2021-2022 were all fully online as opposed to the mixed-modality nature of Summer of AY 2022-2023. In addition, the list of courses offered during the Summer of AY 2021-2022 and AY 2022-2023 were similar to each other. The Summer of AY 2021-2022 ran from June 28, 2021 to August 6,
2021 (LS One Student Blueboard, 2021) while the Summer of AY 2022-2023 ran from June 13, 2022 to July 23, 2022 (Ateneo de Manila University Loyola Schools, 2022). The Canvas logs were first trimmed down so that only the logs that were generated by the 6,519 students and 737 teachers who were present in both Summer of AY 2021-2022 and Summer of AY 2022-2023 are included, and then are tallied overall per date, role type (students or teachers), and class ID, regardless of the event type each log pertains to.

In addition to the Canvas log files, this study also made use of two datasets, one for the Summer of AY 2021-2022 and another for the Summer of AY 2022-2023, containing the list of courses, their sections with their respective Canvas Class IDs, and the learning modality used for the Summer of AY 2022-2023, which includes the following:

- **Fully Online** - All class activities were conducted online during all class days (J. Sugay, personal communication, November 17, 2022).
- **Hybrid** - Some class days were for online class activities, and others were for onsite class activities, for the entire class (J. Sugay, personal communication, November 17, 2022).
- **Flex** - All class activities included students who were onsite and online at the same time during all class days (J. Sugay, personal communication, November 17, 2022).
- **Online+** - Most class activities were online with the exception of special activities (such as exams, presentations, laboratory sessions, etc.) which were onsite (J. Sugay, personal communication, November 17, 2022).
- **Fully Onsite** - All class activities were onsite (face-to-face classes) during all class days.

In order to simplify the assumptions, we only included courses offered during both Summer of AY 2021-2022 and Summer of AY 2022-2023. Furthermore, all sections of any given course offered during Summer of AY 2022-2023 had to follow the same modality.

One of the critical decisions we had to make was the choice of pre-intervention dataset. As mentioned earlier, we had five post-intervention cases: Fully Online, Hybrid, Flex, Online+, and Fully Onsite. The post-intervention dataset for each of these courses was composed of data from all classes offered in these modes. We decided that the pre-intervention dataset of each of these cases would be composed of the same courses offered during the pandemic. To illustrate, suppose that the following were the courses offered during the pandemic and post-pandemic periods as shown in Table 1.

<table>
<thead>
<tr>
<th>Pandemic Courses (All Online)</th>
<th>Post-Pandemic Courses</th>
<th>Post-Pandemic Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A’</td>
<td>Fully Online</td>
</tr>
<tr>
<td>B</td>
<td>B’</td>
<td>Fully Online</td>
</tr>
<tr>
<td>C</td>
<td>C’</td>
<td>Hybrid</td>
</tr>
<tr>
<td>D</td>
<td>D’</td>
<td>Flex</td>
</tr>
<tr>
<td>E</td>
<td>E’</td>
<td>Online+</td>
</tr>
<tr>
<td>F</td>
<td>F’</td>
<td>Fully Onsite</td>
</tr>
<tr>
<td>G</td>
<td>G’</td>
<td>Fully Onsite</td>
</tr>
<tr>
<td>H</td>
<td>H’</td>
<td>Flex</td>
</tr>
<tr>
<td>I</td>
<td>I’</td>
<td>Hybrid</td>
</tr>
<tr>
<td>J</td>
<td>J’</td>
<td>Online+</td>
</tr>
</tbody>
</table>

The pre-intervention and post-intervention datasets would therefore be as shown in Table 2.

<table>
<thead>
<tr>
<th>Post-Pandemic Modality</th>
<th>Pre-Intervention Dataset</th>
<th>Post-Intervention Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Online</td>
<td>A, B</td>
<td>A’, B’</td>
</tr>
<tr>
<td>Hybrid</td>
<td>C, I</td>
<td>C’, I’</td>
</tr>
<tr>
<td>Flex</td>
<td>D, H</td>
<td>D’, H’</td>
</tr>
</tbody>
</table>

Table 1. Courses Offered During the Pandemic and Post-Pandemic Periods (Illustration)

Table 2. Pre-Intervention and Post-Intervention Datasets Based on Table 1 (Illustration)
Note that for this analysis, we did not take into consideration the differences between courses, nor the differences in course contents. The final breakdown of the number of courses and their respective number of sections/classes studied are shown in Table 3.

Table 3. Breakdown of the Number of Courses Analyzed and Their Classes

<table>
<thead>
<tr>
<th>Modality</th>
<th>Total Courses</th>
<th>Total Classes (2021)</th>
<th>Total Classes (2022)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Online</td>
<td>74</td>
<td>126</td>
<td>133</td>
</tr>
<tr>
<td>Hybrid</td>
<td>33</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>Flex</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Online+</td>
<td>9</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Fully Onsite</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

The aggregated Canvas logs and the list of courses were then split according to the academic period. For each academic period, the list of courses was first merged into the aggregated Canvas logs. Then, the resulting dataset was again aggregated according to date, user type, and modality. The total number of classes per modality for the academic period (see Total Classes (2021) and Total Classes (2022) columns in Table 3) was obtained from the list of courses, and was divided from the total number of logs calculated according to date, user type, and modality, and then discretized using a ceiling function to get the average number of logs per class for that date, user type, and modality. Lastly, the t-column represents the number of days elapsed. The values for the Summer period of AY 2021-2022 are computed by subtracting each date d from August 7, 2021, the day after the last day of Summer AY 2021-2022 (LS One Student Blueboard, 2021). The values for the Summer period of AY 2022-2023 are computed by subtracting each date d from June 13, 2022, the first day of Summer AY 2022-2023 (Ateneo de Manila University Loyola Schools, 2022). The t-column therefore has negative values for the Summer period of AY 2021-2022, 0 for June 13, 2022, and positive values for the Summer period of AY 2022-2023. The datasets were then appended back together, then split into separate CSV files (10 in total) according to user type and modality in preparation for CausalImpact analysis using R. Prior to the analysis proper, we take one plus the absolute value of the lowest negative t-column value (that is, pertaining to the first day of Summer period of AY 2021-2022, which is June 28, 2021) (LS One Student Blueboard, 2021), and add it to all of the t-column values in order to conform with the requirement in CausalImpact that t-values need to be positive, continuous, and chronological.

3.2 CausalImpact Analysis

For each modality, CausalImpact was performed with the student data as the outcome variable and the teacher data as the predictor variable. CausalImpact is a type of causal inference that estimates the impact of an intervention (e.g., transition from fully online classes to a variety of learning modalities) using Bayesian structural time series models (Brodersen, 2014; Brodersen et al., 2015; Lagmay & Rodrigo, 2022). Furthermore, in contrast to some other popular causal inference methods like Interrupted Time Series or Prophet (further explained in (Lagmay & Rodrigo, 2022) alongside other alternatives), CausalImpact offers both better predictions and easier interpretations (Kuromiya et al., 2020), although one limitation of it is that it is “only able to observe the outcomes under the treatment for one time series and under the control for the treatment for another one, but not the potential outcome under control for the former and under treatment for the latter” (Li & Bühlmann, 2020). CausalImpact takes in the predictor and outcome variables, as well as the pre- and post-intervention time segments in order to model the relationship between the predictor and outcome variables using the pre-intervention data, estimate the post-intervention
counterfactual, and give the impact of the intervention by measuring the difference between the counterfactual and the observed post-intervention data (Lagmay & Rodrigo, 2022).

We used the teacher data as the predictor variable following the Teacher Expectancy Effect or Pygmalion Effect which states that teacher expectations have an impact on students’ academic performance (Szumski & Karwowski, 2019). A lengthier discussion for our rationale is available in (Lagmay & Rodrigo, 2022). The period pertaining to Summer AY 2021-2022 was designated as the pre-intervention period, while Summer AY 2022-2023 was the post-intervention period.

4. Results and Discussion

4.1 Basic Statistical and Descriptive Analysis

Figures 1 and 2 show the overall average number of logs per class generated by students and teachers, respectively, per modality type and time period, in order to illustrate the difference in LMS usage behavior by user type and modality. For the students data, only the classes using the Fully Onsite modality experienced an increase in the average number of Canvas logs per class (95.54%). This is despite the fact that average teacher activity for the same classes decreased by -29.62%. In contrast, teacher activity in the Flex and Online+ classes saw an increase of 4.93% and 47.88%, respectively.

4.2 Fully Online Classes (Control Group)

Figure 3 shows that the average number of logs per class for Fully Online classes significantly decreased during the Summer of AY 2022-2023 with \( p = 0.014 \). An explanation of the graphs shown in Figures 3 to 7 is as follows:

“Each unit on the x-axis represents one day in the time period. The topmost graph labeled “original” shows a solid line representing the actual observed data, i.e., the number of transactions per day. The broken line represents the prediction. The light blue band represents the confidence interval of the prediction. The middle graph labeled “pointwise”
shows the difference between the predicted number of transactions and the actual number of transactions per day. Finally, the cumulative graph at the bottom shows the accumulated difference between the predicted number of transactions and the actual number of transactions. The [vertical gray dashed line] is the intervention [time]. There is no accumulated difference during the pre-intervention period. The differences are accumulated post-intervention.” (Lagmay & Rodrigo, 2022)

The average value obtained was 98.70 as opposed to the counterfactual prediction of 115.28. The overall estimated effect was -16.58 with a 95% confidence interval of [-31.83, -1.78]. When the data points during the intervention period are summed, the response variable had an overall value of 3,950, as opposed to the counterfactual prediction of 4,610 with a 95% interval of [4,020, 5,220]. Average Canvas logs per class decreased by -14% with a 95% confidence interval of [-24%, -2%].

Figure 3. CausalImpact Graph for Fully Online Classes.

4.3 Hybrid Classes (Experimental Group 1)

Figure 4 shows that the average number of logs per class for Hybrid classes decreased during the Summer of AY 2022-2023, but not significantly. The average value obtained was 102.83 as opposed to the counterfactual prediction of 117.57. The overall estimated effect was -14.74 with a 95% confidence interval of [-35.91, 5.68]. When the data points during the intervention period are summed, the response variable had an overall value of 4,110, as opposed to the counterfactual prediction of 4,700 with a 95% interval of [3,890, 5,550]. Average Canvas logs per class decreased by -12% with a 95% confidence interval of [-26%, +6%]. The difference in the signs meant that the decrease with p = 0.07 was not significant.

Figure 4. CausalImpact Graph for Hybrid Classes.
4.4 Flex Classes (Experimental Group 2)

Figure 5 shows that the average number of logs per class for Flex classes significantly decreased during the Summer of AY 2022-2023 with \( p = 0.001 \). The average value obtained was 149.68 as opposed to the counterfactual prediction of 227.76. The overall estimated effect was -78.09 with a 95% confidence interval of \([-113.40, -43.26]\). When the data points during the intervention period are summed, the response variable had an overall value of 5,990, as opposed to the counterfactual prediction of 9,110 with a 95% interval of \([7,720, 10,520]\). Average Canvas logs per class decreased by -34% with a 95% confidence interval of \([-43\%, -22\%]\).

![Figure 5. CausalImpact Graph for Flex Classes.](image)

4.5 Online+ (Experimental Group 3)

Figure 6 shows that the average number of logs per class for Online+ classes significantly decreased during the Summer of AY 2022-2023 with \( p = 0.001 \). The average value obtained was 69.85 as opposed to the counterfactual prediction of 137.58. The overall estimated effect was -67.73 with a 95% confidence interval of \([-86.15, -51.62]\). When the data points during the intervention period are summed, the response variable had an overall value of 2,790, as opposed to the counterfactual prediction of 5,500 with a 95% interval of \([4,860, 6,240]\). Average Canvas logs per class decreased by -49% with a 95% confidence interval of \([-55\%, -42\%]\).

![Figure 6. CausalImpact Graph for Online+ Classes.](image)

4.6 Fully Onsite Classes (Experimental Group 4)
Figure 7 shows that the average number of logs per class for Fully Onsite classes significantly increased during the Summer of AY 2022-2023 with \( p = 0.001 \). The average value obtained was 122.88 as opposed to the counterfactual prediction of 57.44. The overall estimated effect was 65.43 with a 95% confidence interval of \([49.51, 80.58]\). When the data points during the intervention period are summed, the response variable had an overall value of 4,920, as opposed to the counterfactual prediction of 2,300 with a 95% interval of \([1,690, 2,930]\). Average Canvas logs per class increased by +117% with a 95% confidence interval of [+67%, +191%].

5. Conclusion

We used CausalImpact analysis to examine differences in student LMS participation when all classes were Fully Online versus when they were offered in Fully Onsite, Fully Online, Hybrid, Flex, and Online+ modes. We hypothesized that fully online student online participation would remain unchanged and that participation in all other modes would decline. As hypothesized, online participation in Flex and Online+ classes declined while online participation in Hybrid classes did not differ significantly from activity in pandemic fully online classes. What was surprising though was that participation in Fully Online classes declined while online participation in Fully Onsite classes increased, and these are the findings we would like to unpack further.

From prior literature, we draw several possible explanations for these outcomes. As mentioned in the review of literature, students generally were unsatisfied with the online learning format. They found it isolating and lonely (Singh et al., 2021). They much preferred face-to-face classes or modes that mixed online and onsite formats (Sharma & Alvi, 2021). They suffered “Zoom fatigue” and lacked interactions with peers and mentors (Corral & Fronza, 2022). Working online made them feel less productive (Singh et al., 2021). Furthermore, classes that were purely asynchronous made students feel less capable of understanding course content (Hews et al., 2022). To continue online meant to continue to experience these feelings, all of which may have diminished interest and engagement.

When given the opportunity to work onsite, students felt more productive (Singh et al., 2021). They said that in-person classes with teacher and peers contributed to greater focus and engagement (Hews et al., 2022). They could ask questions as these arose.

The main limitation of this study is that it relies purely on LMS data. It does not have any survey data or qualitative feedback that might help explain the results, and we did not consider the differences among courses nor the learning design utilized by the courses. While references to prior work are useful, the populations across all these studies are different and therefore differences in student experiences may exist. Future work should attempt to collect and analyze more qualitative feedback to weave a fuller narrative explaining pandemic versus post-pandemic student online activity. Also, a more fine-grained analysis such that the logs are aggregated according to event type (e.g. assignments,
quizzes, etc.) should be considered in order to determine which specific parts of the LMS actually increased or decreased depending on the modality.

The other limitation is that the study does not capture the case where students were enrolled in classes with different modalities. Many students had a mix of onsite, online, and mixed-mode classes, and the impact of this intermingling on their overall participation and engagement was not considered in this analysis. In addition, we did not consider the possibility that some students may have failed a course and had to retake it, which may also affect the results. However, such data is not readily available from the Canvas log files alone. Lastly, this study did not take into consideration previous studies on fully online and hybrid courses prior to the pandemic, and as well studies related to the development of digital skills which may help further explain the results.

Despite these limitations, this study contributes to the discourse in at least three ways. It offers an analysis of empirical data from an underrepresented country. It illustrates quantitatively how student participation changed from the pandemic to the post-pandemic period. These changes are a cue to data scientists to think about how to build more robust models that generalize in the fact of major events such as COVID-19. Overall, the findings give educators insight as to what to expect when the next lockdown occurs, and how we might better prepare and recover from it.

Acknowledgements

The authors would like to thank the Ateneo Laboratory for the Learning Sciences and Ateneo de Manila University Research Council for the funding and support needed for this research. The authors would also like to thank Deni Jaramillo and Miguel Saavedra for their assistance in setting up the necessary servers for the collection of Canvas data, and as well Deni Jaramillo and Joannie Ereño for providing the other datasets required for this study. Lastly, the authors would like to thank Jessica Sugay for her explanation on the different learning modalities.

References


Using Augmented Reality to Facilitate Music Learning for Preschool Children

Xiaodong WEI*, Rui QIU* & Ruixue LIU*
*aSchool of Educational Technology, Northwest Normal University, China
*wxd1633@163.com

Abstract: New technologies such as Augmented Reality (AR) have been widely used in education fields. AR’s deep integration of virtual objects and real environments meets the diverse learning needs of learners. However, relatively few studies have explored the application of AR in preschool music education. Therefore, this study developed a mobile AR application for preschool children in learning music and investigated its impact on children’s learning achievement and enjoyment. We used a mixed research method, measuring academic performance through pre-test and post-test, a questionnaire for children’s enjoyment, and interviews to determine children’s views of AR usage in music learning. Sixty participants in two kindergarten classes participated in this study. The children in the experimental group used mobile-based AR applications for music learning, while the control group learned the same content through the traditional teaching method. The results showed that compared with the traditional teaching method, mobile AR has significantly improved children’s academic performance and enjoyment. In addition, the interview results also showed that the children considered AR learning activities to be very interesting and that they are willing to use AR for music learning in the future. The conclusions provide a concrete practical case on how to effectively integrate AR into preschool music education.

Keywords: Augmented reality, preschool children, music learning, academic achievement, enjoyment

1. Introduction

Music is important for promoting children's comprehensive and harmonious development (Klim-Klimaszewska, 2015). Generally, music learning is mainly conducted through visual and auditory activities, and traditional teaching methods (e.g., singing) (Francesconi et al., 2013). However, He (2019) proposed that traditional ways of teaching music are usually the teacher sings first, and then children sing alone, which is unattractive and boring. Previous studies have proposed that Augmented Reality (AR) can be regarded as strategies to improve music teaching levels (Lu et al., 2022). Therefore, incorporating AR in music learning provides the potential to promote learners' effective learning.

AR is a technology that integrates the real environment with the virtual environment which provides learners with an interactive learning experience (Azuma, 1997). AR superimposes designed virtual objects on the real environment through various technologies to achieve the effect of virtual and real fusion (Karagozlu et al., 2019). Cote and Diaz (2017) believed that AR plays an active role in music education, and the AR music teaching materials produced can cultivate students' autonomy. However, few studies explored the application of AR in preschool music education. Therefore, this study aims to integrate AR into musical learning and explore its impact on children's academic performance and enjoyment. The following research questions were as follows:

- Is there a significant difference in the academic performance of children who use AR and the traditional teaching method?
- Is there a significant difference in the enjoyment of children who use AR and the traditional teaching method?
- What are the children’s views on using AR in music learning activities?
2. Background

2.1 Music Learning in Early Childhood Education

Many studies have shown the importance of music learning in the process of early childhood development. Pearce and Rohrmeier (2012) proposed a link between music and cognitive science in their study. The results showed that music is a universal feature of all people, and plays a vital role in our daily lives, as well as in different stages of human development. Qiao (2016) proposed that music education is a practical activity with rich emotions that can promote the emotional development of preschool children. Moreover, music learning can stimulate preschool children’s imagination, improve creativity, activate thinking and intelligence (Klim-Klimasewska, 2015), promote social interaction, cultivate a good sense of social responsibility (Hallam, 2010), and develop positive social values (Ozturk & Can, 2020).

2.2 AR in Music Learning

Some scholars have used AR music systems for music therapy. For example, Correa et al., (2009) showed that AR music learning system can provide therapeutic intervention in cognitive, motor, and psychological aspects to help children with cerebral palsy recover. Correa et al., (2017) showed that the AR music system was beneficial to the rehabilitation of children with disabilities. Moreover, Huang et al. (2011) created an AR-based markerless piano teaching system for piano beginners, their results demonstrated that the AR system increased the sense of accomplishment and interest. Lu et al. (2022) found that using AR to learn chord knowledge in music theory has a positive impact on children’s learning and creativity. The above-mentioned studies demonstrated the advantages of AR in music therapy and education. However, research on the effects of using AR for music learning in early childhood education has not been explored.

3. Development of a Mobile-based AR Music Learning Application

3.1 Learning Material

In this study, we used the Unity game engine to develop an application that can be installed on the Android system. Variety of technologies, such as cube recognition, and multi-image recognition was utilized to develop the AR system. The learning material is the “Forest Kingdom Concert”, which aims to improve children’s singing skills in contextualized learning. There are three main steps, and the specific learning content is shown in Fig 1.

![Figure 1. The Home Page of the Learning Activity.](image)
3.2 Activities for Numbered Musical Notation and Musical Scales

Before learning the numbered musical notation, we selected five different heights of cubes as the identification objects (see Fig 2a). It presented the numbered musical notation of “Do, Re, Mi, Fa, Sol” respectively. When the children scan the yellow-black cube with a mobile phone (see Fig 2b), the 3D virtual model will run out from behind the cube and rotate around the cube, making a “Hello, I’m Fa” sound. Children can control the 3D virtual model through operations (zooming in, zooming out, rotating, etc.). During the musical scale learning process, we used two bar images with different heights and colors to present musical scales: “Do, Re, Mi, Fa, Sol” and “Sol, Fa, Mi, Re, Do” (see Fig 2c). When children scan the musical scale picture, the 3D virtual object is superimposed on the picture, accompanied by the sound of musical scales (see Fig 2d). The two learning activities each last 35 minutes.

![Figure 2. Numbered Musical Notation and Musical Scale Learning Activity Screen.](image)

3.3 Activities for Sing Learning

In the singing learning activity (see Fig 3), the animal’s feet for the 3D virtual object coincide with the footprints of the image to achieve the effect of virtual and real fusion. Children can freely observe and manipulate the 3D virtual object. For example, they can zoom in or out of a 3D virtual object with two fingers and rotate them with one finger (see Fig 3b). After that, the children can click the animal buttons, and click on the animal’s footprint (virtual button) to interact with a 3D virtual object (see Fig 3a). At this time, the lamb’s singing video shown on the picture for children to learn (see Fig 3c). This learning activity lasts 35 minutes.

![Figure 3. Singing Learning Activity Screen.](image)

4. Methods

4.1 Research Design and Participants

In this study, we used a mixed-study method. Before the experiment, both groups of children took a pre-test determining that they had equal prior knowledge levels. Following that, the experimental group used AR to learn music, and the control group learned through traditional methods. After the experiment, a post-test and enjoyment questionnaire was administered to the two groups of children, and a semi-structured interview was conducted. Participants came from two large classes in a kindergarten in northwestern China, each with 30 children, an average age of 6. The children in both groups were taught by the same teacher and were familiar with mobile device.
4.2 Measurement Tools

4.2.1 Learning Achievement Tests

The content of the pre and post-test were the same and was developed by the researcher. The pre-test aimed at assessing whether the children had the same knowledge level before the experiment. The post-test explores whether children who use AR apps for music learning show better academic performance compared to the traditional method. Both pre and post-test questions consist of 10 items, consisting of 5 choice questions, and 5 yes or no items, with a maximum score of 100. For example, the choice question is: “Please select the numbered musical notation of Do”. The reliability of both tests was assessed using Cronbach’s alpha value, which is greater than 0.80, indicating good internal consistency.

4.2.2 Enjoyment Scale

The enjoyment scale was used in this study based on Dalim et al. (2020). It consisted of 7 items (e.g., “I enjoyed doing this activity very much”), assessed using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). To measure happiness, we use a smiley image instead of the Likert scale’s answer categories. The scale is used by preschool children to find emojis that fit their answers with their fingers, and all are filled out by the teacher. The Cronbach’s alpha values for the enjoyment scale were above 0.83, indicating good reliability.

4.2.3 Semi-structured Interviews

The semi-structured interview focuses on understanding preschool children’s views of the AR technology learning experience. In this semi-structured interview, questions include children’s feelings and willingness to use AR applications.

4.3 Procedure

The experiment lasted for about three weeks. In the first week, we pre-tested two groups of children. Moreover, the children in the experimental group were required to be familiar with the operation of AR. In the second week, children in the experimental group learned music through AR, while the control group learned the same material through the traditional method. In the third week, post-test and enjoyment questionnaires were administered to both groups of preschool children. Preschool children in the experimental group also conducted semi-structured interviews.

5. Result

5.1 Learning Achievement

As shown in Table 1, the M and SD of the pre-test scores were 29.93 and 10.26 for the experimental group, and 29.50 and 11.24 for the control group, respectively. The t-test results ($t = 0.16, p > 0.05$) showed that there was no significant difference between the two groups. Therefore, both groups had the same prior knowledge before learning the activity.

After the experiment, the M and SD of the post-test were 87.03 and 6.84 for the experimental group, and 83.50 and 5.38 for the control group, respectively. The t-test results ($t = 2.22, p < 0.05$) showed significant differences between the two groups, indicating that children who use AR applications in music learning showed better academic performance compared to children with the traditional teaching method.
Table 1. T-test Result of the Learning Achievement for Both Groups

<table>
<thead>
<tr>
<th>Test</th>
<th>Group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>F</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental group</td>
<td>30</td>
<td>29.93</td>
<td>10.26</td>
<td>0.42</td>
<td>0.16</td>
</tr>
<tr>
<td>Pre-test</td>
<td>Control group</td>
<td>30</td>
<td>29.50</td>
<td>11.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-test</td>
<td>Experimental group</td>
<td>30</td>
<td>87.03</td>
<td>6.84</td>
<td>1.02</td>
<td>2.22*</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>30</td>
<td>83.50</td>
<td>5.38</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.05.

5.2 Enjoyment

Table 2 shows the t-test results of the enjoyment, the M and SD were 4.22 and 0.57 for the experimental group and 3.85 and 0.69 for the control group, respectively. The t-test results (*t* = 2.28, *p* < 0.05) showed that AR applications for music education had a significant impact on the enjoyment of preschool children.

Table 2. T-test Results of the Enjoyment for Both Groups

<table>
<thead>
<tr>
<th>Item</th>
<th>Group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>F</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental group</td>
<td>30</td>
<td>4.22</td>
<td>0.57</td>
<td>0.91</td>
<td>2.28*</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>30</td>
<td>3.85</td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.05.

5.3 Interview

From the interview results, most children believed that AR has a positive effect on their music learning. Among them, 80.33% of children considered that AR music lessons are very interesting, and learning content makes them feel very happy. 90.33% of children like to use AR for music learning. 82% of children are willing to continue using AR for learning in the future. Here are a few examples taken from interviews with children: “I hope the teacher will let us play AR games next time.” “I like the little animals in AR.”

6. Discussion and Conclusion

In this study, we developed a mobile-based AR application and investigated its effect on preschool children’s music learning. Results found that the children who use AR in music learning achieved better academic performance and enjoyment compared to the traditional teaching method. Moreover, Through the interview, it can be found that the children gave positive feedback on the use of AR in music learning. Regarding achievement, the findings are consistent with previous research findings (Lu et al., 2022). AR increases the fun and provides preschool children with an interactive experience. In addition, in terms of enjoyment, it can also be concluded that the experimental children have higher enjoyment in musical activities than the control group. This is similar to previous research findings (Aladin et al., 2020). In addition, interview results showed that many children believe that AR technology has a positive impact on their music learning, and they enjoy using AR for music learning and look forward to using AR to learn music in the future. AR provides learners with an interactive teaching process and rich learning forms and provides a reference for educators who implement activities in similar fields.

However, this study also has some limitations. First, the children’s influence on the novelty of AR technology may contribute to children’s high satisfaction. Second, some practical activities may be carried out, allowing teachers to lead children to interact with each other. Third, the teaching time of this experiment is relatively short. Future research should enlarge more related knowledge into the music curriculum to form a unified teaching.
Acknowledgments

The research was supported by the National Natural Science Foundation of China [grant number 62067009], and the 2023 Gansu Province University Youth Doctoral Support Project (No. 2023QB-110).

References


Comparing Perceived Cognitive Load while Learning Online with AI Chatbots, Pre-recorded Videos, and Live Lectures

Haixi Sheng$^a$, Xinran Zhou$^b$, Yue Zhao$^c$ & Guqing Zhao$^{d*}$

*a.b.c.d* Faculty of Education, Beijing Normal University, China

*guoqingzh@bnu.edu.cn*

Abstract: The Cognitive Load Theory provides an important perspective for observing learning with different online learning technologies. Since more and more professional development programs are carried out online for in-service teachers, it is crucial to explore their perceived cognitive load brought by different online technologies during online learning. This study aimed to examine the Extraneous Cognitive Load (ECL) and Germane Cognitive Load (GCL) of teachers when learning with three online learning technologies (AI chatbot, pre-recorded videos, and live lecture). Altogether 220 teachers from different areas and schools participated in the study by enrolling on the 21-day online learning training about visual thinking tools. The participants reported they perceived ECL and GCL using a self-reported scale. The results showed that they perceived significantly lower ECL during learning with pre-recorded videos than with AI chatbots and live lectures. Meanwhile, no significant differences appeared in the GCL among these three online learning technologies.

Keywords: Online Learning, Professional Development, ECL, GCL, AI Chatbot

1. Introduction

Technology has changed how we learn and teach. Initially, it was the pre-recorded video lecture represented by MOOC domain the highland of online learning. During the COVID-19 pandemic, live lectures represented by Tencent Meeting and ZOOM have become the mainstream form of teaching. Recently, there is a new way of teaching using AI chatbots, which are like talking robots. However, different online learning methods have different effects on learners. This study looks at how these methods affect teachers’ thinking and learning. To understand this effect, we employ the Cognitive Load Theory (CLT).

The CLT theory says that when we learn something, our brain has to deal with three kinds of load: Intrinsic, Extrinsic, and Germane (F. Paas et al., 2003). Intrinsic load is how hard the content is. Extrinsic load is how hard the method is. Germane load is how well we can use what we learn. The theory says that we should try to lower the extraneous load and increase the germane load to learn better (Sweller et al., 2019). Many researchers have used this theory to study online learning problems (Anmarkrud et al., 2019).

Online learning methods give teachers more choices, but they also need more skills. Teachers have to know how to use different methods for different situations, develop engaging and effective online activities and manage their own thinking and learning with the help of technological tools. This study measures how live lectures, pre-recorded videos, and AI chatbots affect teachers’ extraneous and germane load.
2. Literature Review

2.1 Online Learning and Technologies

Distance education has become popular, and its effectiveness has continuously improved with technological development. Among the existing online learning technologies, pre-recorded videos and live lectures are currently the most common online teaching technologies. The effectiveness of pre-recorded videos and live lectures has been proven in previous research studies (Bertsch et al., 2007; Islam et al., 2020; Shqaidef et al., 2021). Nevertheless, students could not get an opportunity to ask questions and get feedback promptly while learning through pre-recorded videos. With the advancement of technology, live lectures emerged and addressed this weakness. Live lecture mimics face-to-face classes. The instructor presents himself on one side of the computer. And gestures on the other side, ask questions and interact with each other through the chat function. However, the live lecture is limited by technology, Internet speed, and live broadcast equipment (Chitanana et al., 2008). As the latest emerging technology of online learning and instruction, AI chatbot can assist the instructor converts the instructional design into a dialogue in advance, and the learning occurs in a question-and-answer interactive format. AI chatbot can not only embed images, videos, tests, and other content but also collect learning data. Meanwhile, AI chatbots can enhance interactive learning. In a survey on the application of AI chatbots in India, most students stated that AI chatbots can help them solve their problems most quickly (Sandu & Gide, 2019). Although the AI chatbot allows for timely interaction, it lacks emotional transfer. There is a lack of research that compares AI chatbots, pre-recorded videos, and live lectures. Further research is needed to explore whether there are differences in online learning.

2.2 Cognitive Load Theory and its Use in Online Learning

Cognitive load refers to the total mental activity imposed on working memory at any given moment when performing a specific task (F. Paas, 1992). Sweller (1988) first proposed the Cognitive Load Theory (CLT) in the process of problem-solving. Meanwhile, Sweller (2019) divided cognitive load into three categories: Intrinsic Cognitive Load (ICL), Extrinsic Cognitive Load (ECL) and Germane Cognitive Load (GCL). Effective instructional design should minimize ECL as much as possible while using the same ICL to set more cognitive resources for GCL (Sweller et al., 2019).

The three types of cognitive load has also been widely discussed. The definition of the three types of cognitive load shows that the complexity of content determines intrinsic cognitive Load (ICL). At the same time, online learning technology is a vehicle of knowledge for its concerns with the presentation of learning materials and the processing of knowledge by learners. Online learning technology could not influence ICL. Based on the previous cognitive load model (Sweller, 1988), Skulmowski and Xu (Skulmowski & Xu, 2022) proposed the Cost-benefit model in digital learning, which can be seen in Figure 1. In the Cost-benefit model, ECL is seen as a cost, while GCL is seen as a benefit. ECL is considered the effort required for learners to understand the presentation of learning materials, while GCL is the extent to which learners utilize their working memory resources. The Cost-benefit model supports ECL and GCL as essential considerations in online learning.

![Fig 1. The evolution of the models of cognitive Load (Skulmowski & Xu, 2022)](image-url)
2.3 The Measurement of Cognitive Load

Objective and subjective measurements are currently the main ways of measuring cognitive load (F. Paas et al., 2003). Subjective measurement is participants who report their cognitive load by Self-reported. Objective measurement can be psychological, physiological, procedures of tasks, time used in tasks, or reaction time. Objective measurements may cause different results in a survey for its sensitivity and physiological quality. Additionally, objective measurement relies on expensive equipment, and task performance is highly related to the task type. Therefore, subjective measurement may be better than objective measurement in this research study.

There are many self-reported scales for the cognitive load. In the early 1990s, Fred Paas (1992) developed the first cognitive load scale. This Uni-dimensional 9-Point Symmetrical Category Rating Scale has excellent consistency and is widely used in research (F. Paas et al., 2003). Nevertheless, these existing scales do not notice categories of cognitive load. Leppink et al.( 2013) developed a scale to measure different types of cognitive load, which contains the extraneous and intrinsic cognitive load. In this study, we adapt the Cognitive Load Scale (Leppink et al., 2013) for data collection.

2.4 The Present Study

In sum, different online learning technologies may bring different ECLs to learners and allow different GCLs, which effectively indicates learning involvement. This study aims to investigate ECL and GCL perceived by in-service teachers while learning online with three online learning technologies (AI chatbot, pre-recorded videos, and live lecture). The research questions are as follows:

- What are the in-service teachers’ perceived ECL and GCL while learning online with three online learning technologies?
- How do the in-service teachers’ perceived ECL and GCL differ among these three learning technologies?

3. METHODOLOGY

3.1 Participants

This online training lasted 21 days, starting on July 10, 2022, and ending on July 31, 2022. A total of 399 teachers participated in this course. After the training, the teachers who attended the online training completed questionnaires on the widely used online questionnaire platform (https://www.wjx.cn/). We collected 220 valid questionnaires with a valid return rate of 55.10%. All participants voluntarily completed the anonymous questionnaire, and their privacy was ensured not to be disclosed. 15.45% (n=34) of them are males and 85.55% (n=186) are females. Teachers are involved in all subjects. 64.09% (n=141) of the teachers came from primary schools, 24.55% (n=54) from junior high schools, and 11.36% (n=25) from senior high schools.

3.2 Materials

The main aim of this study was to explore online learning technologies which were more effective between AI chatbot, pre-recorded videos, and live lectures. To compare the difference in GCL perceived by teachers when using different online learning technologies, we arranged a 21-day online training and applied an online instructional platform to provide the participators with these three technologies.

To investigate the impact of different online learning technologies on ECL and GCL during learning, we adapted a questionnaire based on the Cognitive Load Scale (Leppink et al., 2013). The adapted Cognitive Load Scale effectively evaluates the ECL and GCL perceived
by teachers, including 13 items about basic information, ECL, and GCL. Moreover, no personal information that could reveal the identity of the participants was gathered.

3.3 Procedure

All 220 teachers who completed a validated questionnaire participated in a 21-day online training course encompassing understanding, manipulating, applying, and evaluating visual thinking tools. The learning was divided into ten modules using an AI chatbot, pre-recorded videos, and live lectures. The learning process was explained as follows: Initially, teachers learned the concepts and cases through AI chatbot and pre-recorded videos and then participated in live lectures to review and consolidated what they have learned in the modules. The live lecture facilitated face-to-face and online communication between the learner and the lecturer, with the lecturer preparing the content beforehand and allowing the learner to ask questions at any time. After the course, a questionnaire was distributed to collect the data needed for the experiment.

3.4 Data Analysis

We used SPSS 24.0 to analyze the quantitative data collected by the questionnaire. Firstly, we examined the reliability and validity of the measurement questionnaire through exploratory factor analysis. Subsequently, one-way repeated-measures ANOVA was used to compare the size of ECL and GCL generated by teachers while learning knowledge using different online learning technologies. Compared to traditional ANOVA, repeated measures ANOVA is more suitable for analyzing the effects of implementing an intervention on multiple variables within the same group of subjects.

4. RESULTS

4.1 Analysis of Reliability and Validity

The reliability and validity of the questionnaire were examined from the valid data. As shown in Table 1, the ECL and GCL scales demonstrated high internal consistency with Cronbach’s α coefficients of .89 and .96, respectively. This questionnaire was adapted from a well-established scale in previous studies, and the scale’s validity was further confirmed by exploratory factor analysis, with KMO=0.89 > 0.7.

<table>
<thead>
<tr>
<th>Factor and Items</th>
<th>Factor Loading</th>
<th>Mean</th>
<th>SD</th>
<th>Cronbach’s α</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ECL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live lecture</td>
<td>.89</td>
<td>2.21</td>
<td>.82</td>
<td>.89</td>
</tr>
<tr>
<td>Pre-recorded</td>
<td>.94</td>
<td>2.24</td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td>videos</td>
<td></td>
<td>2.15</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>AI chatbot</td>
<td>.85</td>
<td>2.25</td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td><strong>GCL</strong></td>
<td>.84</td>
<td>4.13</td>
<td>.53</td>
<td>.96</td>
</tr>
<tr>
<td>Live lecture</td>
<td></td>
<td>4.13</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>Pre-recorded</td>
<td>.88</td>
<td>4.13</td>
<td>.32</td>
<td></td>
</tr>
<tr>
<td>videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2 Analysis of the Effect of Different Online Learning Technologies on In-service Teachers’ ECL and GCL

The results of the repeated measures ANOVA showed that there is a significant borderline effect among ECL of different online learning technologies ($F(1.67,364.73) = 2.708, p = .078$
< .10, partial $\eta^2 = .012$), which suggests that teachers produce different ECL when using various technological tools to learn knowledge. ECL perceived by teachers when learning with pre-recorded videos is significantly lower than the ECL of AI chatbot ($p = .016 < 0.05$) and ECL of live lecture ($p = .035 < 0.05$). ECL of the AI chatbot and live lecture have no significant difference ($p = .85 > 0.05$). Regarding the analysis of GCL, we used Greenhouse-Geisser estimates of sphericity ($\epsilon = .95$) to correct degrees of freedom for sphericity violations (Mauchly’s $W = .95, \chi^2(2) = 11.93, p = .003$). Repeated-measures ANOVA showed no significant difference in the GCL when teachers used different online learning technologies, with $p = .59 > 0.05$.

### Table 2. Repeated-measures ANOVA and post hoc tests

<table>
<thead>
<tr>
<th>Variable</th>
<th>F</th>
<th>Sig.</th>
<th>partial $\eta^2$</th>
<th>Post hoc comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECL of different Online Learning</td>
<td>2.708</td>
<td>.078</td>
<td>.012</td>
<td>pre-recorded videos $&lt;$\text{Al}$ chatbot, pre-recorded videos $&lt;$\text{live}$ lecture</td>
</tr>
<tr>
<td>GCL of different Online Learning</td>
<td>.505</td>
<td>.594</td>
<td>.002</td>
<td></td>
</tr>
</tbody>
</table>

### 5. DISCUSSION

This study tries to compare how teachers feel when they learn online with chatbots, videos, and live lectures. The result shows that pre-recorded videos are more effective than AI chatbot and live lectures in reducing ECL. In contrast, AI chatbot and live lectures have no difference. This may be because of some reasons. First, new technologies can make learners stressed and confused (Riedl, 2012). Learners have a better understanding of videos compared to live lectures and chatbots. Additionally, new technologies can encounter errors or teaching-related issues. For example, chatbots may not provide accurate or clear answers, while live lectures can be affected by network problems. As a result, these factors can cause learners to lose their focus or interest. However, it is worth noting that chatbots did not create more confusion for learners compared to live lectures, which indicates that teachers found them beneficial.

None of the three online learning technologies greatly impacted GCL, indicating that extrinsic technology had a limited effect on fostering intrinsic cognitive structures and motivation. Learners need to talk and interact more in the class to understand better. Other studies also said that videos work better with different kinds of knowledge and teacher help (Hong et al., 2018).

This study has some limits. It only asked learners how they felt, which may not show how they really thought. It also did not look at other things that can affect thinking and learning, like what learners know before, how they feel, what they like, or what they want.

### 6. CONCLUSIONS AND IMPLICATIONS

These findings suggest that teachers who learned through pre-recorded videos had a lower perception of their ECL compared to those who used AI chatbots or attended live lectures. In regards to GCL, the three different online learning technologies do not show a significant difference in GCL. This study gives some ideas for future research studies. First, future research studies could utilize more advanced AI technologies to improve the design and implementation of AI chatbot, such as natural language processing models, speech recognition systems, and generative AI. Second, future research studies could expand from teachers to students. We can analyze how different online learning methods affect different groups of learners, such as age, gender, knowledge, and preferences. Third, subjective and objective measurements can be combined to enhance the results from multiple perspectives when measuring cognitive load or considering other relevant variables in online learning will
be a valuable topic. Fourth, future research could focus more on technology-assisted instruction than technology-replaced instruction. Technology can make teaching easier, but it cannot do what teachers do. Teachers do not only give knowledge but also show thinking, skills, and values to learners.

References


https://www.semanticscholar.org/paper/The-current-state-of-e-learning-at-universities-in-Chitana-na-Makaza/70ad52e2263d8fdd037d3d80c8486f4a74cd30


Abstract: In this paper, we replicated the methods and findings of a study on successful completion of massive open online courses (MOOCs) in the context of online learning in a university in Metro Manila, Philippines. Specifically, we examine whether discussion forum behaviors associated with completion in MOOCs also resulted in better outcomes in university online learning courses. To this end, we collected logs from the university’s learning management system. The Mann-Whitney U test was used to compare the average scores of those who conformed to these behaviors and those who did not. Pearson’s r correlation was then used to determine if a significant relationship exists between each of the measured behaviors and the average scores of the students for each course. The same analyses were then performed for each of the five schools in the same university. The study failed to replicate the results of the MOOC study in all cases. The difference in means was not far apart for all rules. This could be attributed to the difference in pacing and demographic between universities and MOOCs. Despite these, the study still demonstrates that techniques used to study MOOCs can be used to study the behaviors and success indicators of students.

Keywords: MOOC, university online learning, academic success

1. Introduction

Massively Open Online Courses (MOOCs) are Internet-based classes offered at minimal to no cost. They are available to any participant with Internet access, anytime, anywhere (Cagiltay, Cagiltay, & Celik, 2020). Because barriers to entry are low, millions of people enroll in these courses. The top 5 MOOC providers—Coursera, edX, XuetangX, Udacity, and FutureLearn—account for a collective population of 90 million learners.

Completion rates for these MOOCs, though, tend to be low, typically between 5 and 15% (Ahern, 2018). Researchers have therefore tried to study the relationship between student in-course behaviors and their course completion. Typically, these studies examined students within a single course, raising questions about replicability. Which of these findings generalize across MOOCs and which do not? Andres and colleagues (2018) addressed the issue of replicability by testing prior findings against data from multiple MOOCs. They found that students who were likely to complete a course tended to spend more time in forums and on assignments, make longer posts, post, and respond more frequently, and so on.

The question that we pose in this paper is whether these findings transfer beyond MOOCs, to online learning during the COVID-19 pandemic. Transferring these findings could provide an alternative method of studying success indicators beyond the usual surveys done for online learning. This study investigates the following behaviors associated with earning a certificate in MOOCs studied by Andres et al. (2017):

- Rule 1. Number of posts in discussion forums is greater than average.
- Rule 2. Length of posts in discussion forums is longer than average.
- Rule 3. Number of responses to others in discussion forums is greater than average.
- Rule 4. Number of respondents to threads in discussion forums is greater than average.
More specifically, the study aims to answer the following research questions:

- **RQ1:** Is there a significant difference between the average scores of students that fall under the behaviors listed and those that do not?
- **RQ2:** Is there a correlation between the average word count per discussion post, the average number of posts, the average number of replies received, and the average number of replies to others, and the average score of each student?
- **RQ3:** Are these correlations present within the courses offered by the different schools of the university?

### 2. Background Literature

#### 2.1 MOOCs and Universities

MOOCs have often been compared to on-campus university courses, especially given that some universities also offer MOOCs. On-campus university courses consist of students who are closer to one another while MOOCs have a wider reach across different countries. University courses also follow a directed learning curriculum with a pre-defined course outline. MOOCs mostly follow self-directed learning with a more flexible timeline. (Begiu & Strobl, 2015). Even educators have a difference in their approach and roles between the two. Educators in MOOCs those courses cited managing the course, creating resources, and guiding students as part of their roles. On the other hand, on-campus university educators cited that guiding and accompanying students formed the bulk of their work while providing feedback carried a lesser weight (Cabrera & Fernández-Ferrer, 2017).

#### 2.2 Student Success in MOOCs and University Online Learning

Research into MOOCs has focused on behaviors associated with completing the course as most MOOCs have a low completion rate (Crossley et al., 2016). The present study focuses on data collected from discussion forums which have similarly been the subject of several studies into MOOC completion. Andres et al. (2017) investigated 21 findings or rules from previous studies to identify if these held true for their dataset. Out of the 21 rules, 14 involved behaviors in discussion forums, but the present study focuses on the four rules mentioned earlier. Out of the four, the study failed to replicate the second rule since the analysis was not statistically significant.

On the other hand, studies into the performance of university students in online learning primarily focus on survey-based questions. A survey of undergraduate nursing students in South Korea by Kim et al. (2022) found that self-directed learning behavior and satisfaction with their course were found to be predictors of academic achievement. Other studies (Basith et al., 2020; Gopal et al., 2021) similarly made use of surveys and explored the relationship between online learning satisfaction and academic performance. It is evident that previous studies have focused on data collection through surveys. As such, the present study aims to investigate the use of techniques in MOOC researchers to understand the behaviors associated with academic success in university online learning.

### 3. Methods

#### 3.1 Description and Preparation of Dataset

The study made use of students’ log data from the Canvas LMS of a university in Metro Manila, Philippines, collected during one semester from August 25, 2021, to December 18, 2021. During this time, all classes at the university were held online. There were a total of 3,429 classes offered and 6,439 students enrolled in the university. The university was divided into 5 schools: education, humanities, management, science and engineering, and social sciences. Many courses within each school can be taken by students from any field of specialization, i.e., a student specializing in management can take a social science course,
and a student specializing in science can take a humanities course. The discussion forum logs from each course were used to compute the average word count, average number of replies to others, and average number of replies received for each of the students.

In the study of Andres and colleagues (2017), success was defined as the earning of a certificate. In our dataset, though, students earned grades that ranged from 0 to 100, so success had more variation. Hence, we used the grade logs for all graded assignments and computed the average score percentage of each student for each course.

3.2 Comparison of Means for each Rule

Since the success indicator (average score across all graded assignments) is no longer binary (completed or did not complete the course), a comparison of means was used instead of the chi-square test used by Andres et al. (2017). For each course, a course average for each of the four rules was obtained to split the students into two groups for each rule: those who are above average and those who are below average. A test was then used to compare the means of each group for all four rules. In choosing the test, the normality of the data was first assessed using Q-Q plots (Das & Imon, 2016). Q-Q analysis for all four rules showed that all four data are not normal. Thus, the Mann-Whitney U test (McKnight & Najab, 2010) was used to test if there is a significant difference in the average score percentage of those who are above average and those who are below average for each rule. The same analysis was done for the data after being divided by school.

3.3 Analysis using Correlation for each Rule

Given that both the rules and the success indicator are continuous variables, it is possible to identify if there exists a correlation between each rule and the average score percentages. For this, Person’s r correlation (Akoglu, 2018) was used. This analysis provides more insights for courses in the university where academic success is dependent on numerical grades and not a binary variable (pass or fail).

4. Results and Discussion

4.1 Analysis Across the Entire Population

Table 1 presents the median and interquartile range for the two groups under each rule. The results of the Mann-Whitney U test show that there is a significant difference (p≤0.05) between the mean score percentage of the above and below-average groups for Rules 2–4. This seems to follow the findings of Andres et al. (2017), but the median and interquartile range of the average score percentage per group doesn’t support the replication. The two median average score percentages of the two groups under Rules 2–4 only differ by at most 1%. Similarly, the interquartile ranges of both groups are mostly overlapping.

Table 1. Median and IQR for the average score percentage of the groups in each rule.

<table>
<thead>
<tr>
<th>#</th>
<th>Rule</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Average number of posts</td>
<td>Above average</td>
<td>91.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Below average</td>
<td>90.67</td>
</tr>
<tr>
<td>2</td>
<td>Average word count*</td>
<td>Above average</td>
<td>91.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Below average</td>
<td>90.00</td>
</tr>
<tr>
<td>3</td>
<td>Average number of replies given*</td>
<td>Above average</td>
<td>91.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Below average</td>
<td>90.65</td>
</tr>
<tr>
<td>4</td>
<td>Average number of replies received*</td>
<td>Above average</td>
<td>91.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Below average</td>
<td>90.63</td>
</tr>
</tbody>
</table>

*statistically significant (p ≤ 0.05)
The results of the Pearson’s $r$ Correlation in Table 2 provide more insights into the data. Only the correlation coefficients for Rules 1 and 2 were found to be statistically significant ($p \leq 0.05$). For Rule 1, the result shows that there exists a negligible correlation between the students’ average number of posts and their average score percentage. Similarly, for Rule 2, the result shows that there exists a negligible correlation between the students’ average word count per post and their average score percentage. These results show that students with a higher average number of posts and word counts did not necessarily have higher average scores.

Table 2. Results of Pearson’s $r$ Correlation across all four rules.

<table>
<thead>
<tr>
<th>#</th>
<th>Pearson’s $r$ coefficient</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Average number of posts</td>
<td>-0.041</td>
</tr>
<tr>
<td>2</td>
<td>Average word count</td>
<td>0.060</td>
</tr>
<tr>
<td>3</td>
<td>Average number of replies given</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>Average number of replies received</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*statistically significant ($p \leq 0.05$)

Overall, the results suggest that the study failed to replicate the results of Andres et al. (2017). Although the Mann-Whitney U test showed a statistically significant difference with the means for Rules 2–4, the difference turned out to be at most 1%. For Rule 2, this point was confirmed by the Person’s $r$ Correlation which showed negligible correlation. Moreover, the Person’s $r$ Correlation also showed that Rule 1 may not hold true as well given the negligible correlation. These results could be explained by how university instructors often give more rigid guidelines—such as requiring only one post—compared to those in MOOCs. Similarly, results show that the length of these responses don’t have a significant relationship with the students’ performance. It might be more insightful, then, to analyze the content of responses. Aside from instructor guidelines following the directed learning approach (Belgiu & Strobl, 2015), the failure to replicate could also be attributed to the difference in demographic between universities and MOOCs.

### 4.2 Analysis per School

Tables 3 and 4 show the results of the same analyses conducted on the dataset divided by the school that offers each course. The results of the Mann-Whitney U test for all schools show that, among those with a statistically significant difference in means, the difference in the medians is at most 1%, and the interquartile ranges are mostly overlapping. The only exceptions are Rules 3 and 4 for the science and engineering courses where a difference of 3% was observed. However, the interquartile ranges for the two groups under both rules are still mostly overlapping. Moreover, the Pearson’s $r$ correlation for Rules 3 and 4 of the science and engineering courses show negligible correlations for both rules. Thus, we are unable to conclude that these courses adhere to Rules 3 and 4. The remaining statistically significant results for the Pearson’s $r$ correlation show negligible. Therefore, even when divided by school, the study fails to replicate the results of Andres et al. (2017), and the results are similar to the results of the analysis done across the entire population. This could indicate that instructors across schools provide the same rigid guidelines described earlier.

Table 3. Median and IQR for the average score percentage of the groups in each rule divided by school.

<table>
<thead>
<tr>
<th>School</th>
<th>Rule</th>
<th>Below Average</th>
<th>Above Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td>IQR</td>
</tr>
<tr>
<td>education</td>
<td>1</td>
<td>96.25</td>
<td>93.45 – 99.57</td>
</tr>
</tbody>
</table>
Table 4. Results of Pearson’s r Correlation across all four rules per school.

<table>
<thead>
<tr>
<th>Rule</th>
<th>education</th>
<th>management</th>
<th>p-value</th>
<th>Pearson’s r coefficient</th>
<th>education</th>
<th>management</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.09878</td>
<td>0.34352</td>
<td>0.05855</td>
<td>-0.09878</td>
<td>0.04901*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.15506</td>
<td>0.13563</td>
<td>0.09572</td>
<td>0.15506</td>
<td>0.00127*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.06646</td>
<td>0.52451</td>
<td>0.06733</td>
<td>-0.06646</td>
<td>0.02354*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.01456</td>
<td>0.88923</td>
<td>0.01896</td>
<td>0.01456</td>
<td>0.52403</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion and Further Studies

Overall, the study failed to replicate the results found by Andres et al. (2017). There are several possible reasons why the MOOC results did not transfer to our population. First, the students enrolled in MOOCs and those enrolled in traditional college courses differ in fundamental ways. MOOC students tend to be between 25 and 65 years old (Cagiltay, Cagiltay, & Celik, 2020). Most already have a higher education degree and are male. They enroll in a MOOC because (1) they want to retool or explore new professional areas, (2) they have a personal interest in the course content, or (3) the course is important for career advancement (Baturay, 2015; Williams, Stafford, Corliss, & Reilly, 2018). Guided by their own motivations, most learners tend to navigate the course in a non-linear fashion, studying what they need and skipping the rest (Cagiltay, Cagiltay, & Celik, 2020). MOOCs enable this type of behavior because students are minimally supervised—it is rare for a student to receive direct and timely feedback from course instructors.

On the other hand, the students in our population were enrolled in traditional courses that were forced online by COVID-19. Hence, students received direct supervision from
teachers, including psycho-social support (see Moreno et al., 2021). Furthermore, culturally, Filipinos place a high value on education (Galang & Reyes, 2009). They see education as a solution to poverty and ignorance, as a means towards better life opportunities. Finally, Filipinos are motivated to persevere in their studies by social cues from parents, teachers, classmates, and friends. These factors may have driven students to succeed in their courses, despite the difficulties brought on by the pandemic.

Although the study was unable to replicate the results, it was still able to demonstrate how techniques used to study MOOCs can be used to study the behaviors and success indicators of students. This deviates from the usual surveys to study student behavior and makes use of data collected from a learning management system. Moreover, future studies can still consider replicating the other rules that the current study was not able to consider.

Acknowledgements

The authors would like to thank the Ateneo Laboratory for the Learning Sciences (ALLS), Ateneo Research Institute for Science and Engineering (ARISE), and Accenture for the funding and support for this research. The authors would also like to thank Deni Jaramillo and Miguel Saavedra for their assistance in setting up the necessary servers for the collection of Canvas data, Ezekiel Lagmay for his assistance with extracting and cleaning the data, and Francine Tarrazona for proof-reading the paper.

References


Visual Attention Patterns in Processing Compiler Error Messages

Christine Lourrine TABLATIN\textsuperscript{a,b} & Maria Mercedes RODRIGO\textsuperscript{a}
\textsuperscript{a}Ateneo de Manila University, Philippines
\textsuperscript{b}Pangasinan State University, Philippines
*tablatinchristine@gmail.com

Abstract: The difference in the visual attention of subjects while performing a debugging task can be measured using fixation count and fixation duration metrics. Thus, this paper investigated the visual attention patterns of high and low performing students engaged in a defect-finding task on multiple programs using these metrics. We performed statistical tests on the proportional fixation durations and fixation counts on the error lines and the compiler error messages to determine the difference in the visual attention patterns between the groups. The results of the statistical analysis revealed a significant difference between the high and low performing students across all programs. This implies that high performing students were associated with significantly higher visual attention on the error lines of the programs than the low performing students. However, the analysis of the proportional fixation duration and fixation count on the compiler error messages revealed no significant difference between the groups. The results suggest that both groups showed similar visual attention to the compiler error messages. The findings of this study provide insights into the visual attention patterns of student programmers in processing compiler error messages. High and low performing students could be distinguished based on their visual attention patterns on the error lines but not on the compiler error messages. Further, high performing students prefer a more analytical processing approach and pay attention to relevant code elements in debugging to correctly identify the errors while low performing students choose a more holistic approach.

Keywords: visual attention, compiler error messages, debugging, eye-tracking, fixation metrics

1. Introduction

Debugging is a critical skill that remains challenging to acquire for student programmers. Students rely on compiler error messages to aid them in fixing the errors in their programs while debugging them. Compiler error messages are pedagogically significant to student programmers since they give feedback on what went wrong in their programs (Becker et al., 2019). However, according to Du Boulay & Matthew (1984), computer programming students are unable to relate compiler error messages to actual code errors. One of the possible reasons for this is that compilers are well-known for generating cryptic and uninformative error messages (Barik et al., 2017). This problem still exists and contributes to a high attrition rate of students in computer programming (Becker et al., 2018). Therefore, researchers invested efforts in developing systems that offer students more insightful error messages (Nienaltowski et al., 2008; Dy & Rodrigo, 2010; Denny et al., 2014; Prather et al., 2017). The work of Nienaltowski and colleagues (2008) found that providing additional information regarding an error did not necessarily result in greater debugging accuracy. Denny et al. (2014) corroborated their finding. In contrast, the study of Prather, et al. (2017) found that enhanced compiler error messages are more helpful than standard compiler error messages through an eye tracking study.

By observing people's eye movements, researchers can determine what attracts them and possibly understand how information is interpreted (Bol et al., 2017). As a result, researchers have used eye tracking to measure the difficulty student programmers
encounter in understanding the error messages they read. Barik et al. (2017) found that interpreting error messages is as challenging as reading source code. Some of the reasons comes from the lack of familiarity with error messages (Marceau et al., 2011) and the need for programmers to move between the error message and source code to understand the full context of the problem. Other researchers have examined the reading patterns that students employ to parse through code to find an error and how these patterns differed depending on student ability. The source codes used in some studies were written in C (Chandrika & Amudha, 2017; Nivala et al., 2016; Sharif, Falcone & Maletic, 2012), C++ and Python (Turner et al., 2014), and Java programming languages (Villamor & Rodrigo, 2022; Tablatin & Rodrigo, 2022). Although research on understanding how programmers read and comprehend source code has been conducted, research on processing compiler error messages is still limited, especially using C++ as the source code stimuli. Through an eye-tracking study, we hope to add to what is known about how student programmers read and process compiler error messages. In this paper, we investigated the visual attention patterns of high and low performing students in processing compiler error messages.

2. Methodology

2.1 Participants

The participants of this study were Information Technology students who have at least taken a college-level introductory programming course using C++ as the programming language. A total of 32 students aged 18 years old and above were recruited from a state university in the Philippines. The study used two participant groups: high performing and low performing. The scores of the participants in the debugging tasks were used to assign them to a particular group. High performing group consisted of students who scored above and equal to the mean score while the low performing group consisted of students who scored lower than the mean score.

2.2 Experimental Setup and Procedure

All participants underwent a screening process to determine their eligibility to take part in the study. Students who passed the initial screening were given an informed consent form to fill out and sign. The students who signed the informed consent forms were given two types of pre-tests: a personality test and a self-efficacy test. After the pre-tests, the participants were required to undergo a calibration test before starting with the eye-tracking experiment that was designed to last for about 30 minutes.

The experiment was conducted in a laboratory setting. The experimental setup consisted of a laptop, a 17-inch monitor, a mouse, and a Gazepoint GP3 table-mounted eye tracker with a sampling rate of 60 Hz and an accuracy range of 0.5–1 degree. The hardware was set up by extending the laptop’s display to the monitor connected to it. The eye tracker was placed in front of the monitor to allow the participant to view the program codes while the eye tracker records the participant’s eye movements. The participant sat in front of the monitor, eye tracker, and mouse. After successful calibration of the eye tracker, a custom-built slide viewer loaded with C++ programs was activated. The slide viewer enabled the participant to navigate through the programs using Previous and Next buttons. It also enabled the participant to mark error locations with a red ellipse using the mouse. A Reset button can be used to clear the marked errors on a given slide and a Finish button saves the participant’s answers and ends the session.

2.3 Comprehension Task and Measures

The main task of the participants was to find the errors of the programs based on the compiler error message. Each program had been injected with one syntax error and there is no need to correct them. Participant performance was measured based on the correctness
of the answers. Eye tracking metrics were also collected to measure the visual attention patterns of the students while performing the debugging task.

2.4 Stimuli Characteristics

The code pool size was limited to 5 unique programs with different syntax errors. The programs shown to the participants were brief (15 to 35 lines) and represent C++ constructs discussed during the first 6 to 8 weeks of a first collegiate programming course. These include data types, variables, operations, conditionals, and loops. Further, the level of difficulty of the code was limited to the types of exercises given to students who are being introduced to the constructs mentioned herein. We have limited the error types to syntax errors that are usually encountered by students who are trying to complete these programming exercises. Only 1 syntax error was injected in each program and no semantic or logical errors were included. We have chosen 5 syntax errors from the syntax errors identified in the study of Denny, et al. (2014) to be injected in the programs. Note that, despite this limitation, some of the errors may be non-literal. That is, they may not accurately reflect the error or its location. The programs used in the experiment are novice-friendly, self-contained, and did not require any special domain knowledge. The language used was familiar to the students, brief enough to fit on a single screen, and did not use unfamiliar APIs. The stimuli used in the experiment were static. Participants only have to point to the location of the error in each program. They do not need to correct it.

2.5 Data Pre-processing

The gaze movements of the participants were stored in a CSV file format. The time of the recording (timestamp) when fixations occur, the location of the fixations (values of x and y coordinates), and the fixation duration of each fixation were extracted from the CSV file for the visual attention patterns analysis. Areas of Interests (AOIs) of the 5 programs were drawn using the OGAMA Areas of Interest module (Vosskühler, 2009) to get the AOI coordinates. The AOIs marked in this study are the line where the error is located and the compiler error message of each program. Figure 3 shows the AOIs of one of the programs used in this study. The AOI coordinates extracted from OGAMA returns coordinates with respect to the setting of the screen resolution specified when the AOIs were defined. To map the location of fixations to the program codes, the x and y coordinates from the eye-tracking data were converted by multiplying the x coordinates with 1366 and the y coordinates with 768. This was done to match the coordinates of the program codes during the experiment since the screen resolution used was 1366 x 768. In addition, the fixation durations were recorded in terms of seconds by the eye tracker and were converted into milliseconds by multiplying the duration with 1000. These processes were done for the eye-tracking data of the 32 students to determine their visual attention patterns.

2.6 Data Analysis

To determine the difference in the visual attention patterns of high performing and low performing students, fixation count and fixation duration were used. The proportional fixation count and proportional fixation duration on the error lines and the proportional fixation count and proportional fixation duration on the compiler error messages were computed to measure the visual effort exerted by high performing and low performing students.

Statistical analysis was conducted to compare the visual efforts between the high and low performing students using the visual effort metrics stated above. Independent samples t-test was used to determine whether there is statistical evidence that the visual efforts are significantly different between high and low performing students.
3. Results and Discussion

The eye tracking data of 32 students from a state university in the Philippines were used in the analysis. Eighteen (18) students were identified as high performing while fourteen (14) students were considered low performing based on their debugging scores. The proportional fixation duration and count on the identified AOIs were computed as a ratio of fixation duration and fixation count on an AOI to the overall fixation durations and counts on each program, respectively. These data were used to determine the difference in the visual attention patterns of students while finding errors in the program.

3.1 Visual Attention Patterns on the Error Lines

To determine the difference in the visual attention patterns of the groups in the error lines, independent samples t-tests were used. An independent samples t-test was used to determine if there is a significant difference in the average proportional fixation duration of the high and low performing students. The result of the analysis revealed that there is a significant difference in the average proportional fixation duration of high performing students ($M = 0.014, SD = 0.007$) and low performing students ($M = 0.007, SD = 0.002$), $t(17.823) = -3.738, p = .002$. The result suggests that the average proportional fixation durations of the high performing students are significantly higher than the low performing students across all programs. Further, an independent samples t-test was also performed to determine if there is a significant difference in the visual attention of high and low performing students in terms of the average proportional fixation count on the error lines. The result of the analysis revealed that there is a significant difference in the average proportional fixation count of the high performing students ($M = 0.013, SD = 0.006$) and low performing students ($M = 0.007, SD = 0.002$), $t(16.411) = -4.165, p = .001$. The result suggests that the average proportional fixation count of the high performing students is significantly higher than the low performing students across all programs.

The results of the analysis of the visual attention patterns suggest that high performing students were associated with significantly higher proportional fixation durations and fixation counts on the error lines of the programs than the low performing students. Fixation durations are influenced by the complexity and difficulty of the visual content, task being performed (Bylinski, et al., 2015; Nivala, et al., 2016), and AOIs that are engaging the cognitive resources of the observer (Bylinski, et al., 2015). High performing students have exerted more visual attention in terms of the proportional fixation duration on the error line to confirm the compiler error message that error does exist in that line. This finding corroborates the findings of Chandrika & Amudha (2017) and Sharif et al. (2012) that experts tend to concentrate more on areas where the errors are located while novices read the codes more broadly. In contrast with the findings of the latter, Nivala, et al. (2016) and Turner, et al. (2014) observed that novices spent more time on the buggy lines of code than experts. The difference in the findings of this study with previous studies may be related to the characteristics of the stimuli used and the tasks employed by the researchers during the eye tracking experiment. Furthermore, the number of fixations on an AOI can be linked to its importance (Bylinski, et al., 2015). Thus, more fixation counts can be observed from the high performing students on the error lines. This result is in line with the findings of (Chandrika & Amudha, 2017; Sharif, et al., 2012; Turner, et al., 2014), that experts or advanced programmers have more fixation counts on the buggy lines of code.

3.2 Visual Attention Patterns on the Compiler Error Messages

An independent samples t-test was performed to determine the difference of the visual attention of high and low performing students on the compiler error messages in terms of the average proportional fixation duration. The result of the statistical analysis revealed that there was no significant difference in the visual attention of the high ($M = 0.023, SD = 0.016$) and low performing students ($M = 0.018, SD = 0.024$), $t(16.066) = -0.696, p = .497$. The result suggests that the average proportional fixation duration on the compiler error
messages of the high and low performing students is similar for both groups across all programs. The same statistical test was used to determine the difference of the visual attention of the groups on the compiler error messages in terms of the average proportional fixation count. The result revealed that there was no significant difference in the visual attention of the high ($M = 0.025$, $SD = 0.015$) and low performing students ($M = 0.018$, $SD = 0.022$), $t(16.256) = -0.926, p = .368$. The result suggests that the average proportional fixation counts is similar for both high and low performing students across all programs.

The results of the analysis of the visual attention patterns imply that both high and low performing students have similar average proportional fixation durations and fixation counts on the compiler error messages. Although no significant difference was observed, high performing students have slightly higher average proportional fixation duration and fixation count on the compiler error messages. The visual behavior of high performing students may be related to the field-independent (FI) cognitive style theory of human cognition. The FI individuals tend to choose a more analytical processing approach and they pay attention to relevant details. Conversely, the visual behavior of low performing students may be related to field-dependent (FD) cognitive style wherein they choose a more holistic way of processing visual information and experience difficulties in identifying relevant details in the complex visual stimulus (Raptis et al., 2017).

4. Conclusion and Future Works

This study contributes to the evidence of the effectiveness of eye tracking as a method to enrich computing education research. The analysis conducted using the proportional fixation duration and fixation count provided considerable insights about the visual attention patterns of high and low performing students in processing compiler error messages. High performing students exert more visual attention on the lines indicated by the compiler error messages. They prefer a more analytical processing approach and pay attention to details while low performing students choose a more holistic approach. We may conclude that more proficient students read error messages as much as the less proficient students but limit their reading to the line number and not to the whole description of the error message. By exploring the visual strategies employed by the high performing students using eye tracking data, we could develop learning materials and activities that could help low performing students improve their code reading and debugging skills. Further, debugging should be taught as a program comprehension task rather than a search task. Students should focus their attention on the lines mentioned in the compiler error message to identify source code defects because it provides feedback on the location and description of the error. Teaching students to consciously employ debugging strategies would enhance their debugging ability and will help in increasing the retention rate of students taking programming courses.

Future analysis on the collected eye tracking data will be conducted to determine if students with different programming ability, self-efficacy levels, and personality characteristics vary in the way they read and process compiler error messages.

Acknowledgements

We thank Ateneo de Manila University and Pangasinan State University administration for allowing us to conduct the experiment for this study. We also thank the Dagstuhl Seminar 22402 on Foundations for a New Perspective of Understanding Programming that enabled us to identify the shortcomings in our previous study design and helped us improved the study design of this paper.

References

May). Do developers read compiler error messages?. In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) (pp. 575-585). IEEE.


Process Evaluation for Concept Map Building and Its Experimental Evaluation

Ridwan RISMANTOa,c*, Aryo PINANDITOa, Banni Satria ANDOKOc, Yusuke HAYASHIA & Tsukasa HIRASHIMAa

aGraduate School of Advanced Science and Technology, Hiroshima University, Japan
bDepartment of Information System, Brawijaya University, Indonesia
cDepartment of Information Technology, State Polytechnic of Malang, Indonesia
*rismanto@polinema.ac.id

Abstract: Concept map as an assessment tool is widely accepted in the field of education. In this research, assuming the process evaluation of concept map building leads to better evaluation than the product (that is, built concept map) evaluation, we proposed a process evaluation method and conducted an experimental evaluation of the method. The process evaluation method is intended for re-composition type of concept maps, not for general ones. In the re-composition concept map, a learner is provided a set of components of a concept map, that is, links and nodes, and required to recompose a concept map by using the components. The provided components are generated by decomposing the concept map built by an expert as a representation of understanding of the learning target. In this research, we set several submaps with a set of semantics in one expert map and define a good process by recomposing the submap as a series of activities. The process evaluation results were compared with the product evaluation results for the data of a practical use of re-composition of concept map in undergraduate class. As the results of the comparison, we have already confirmed that the result of process evaluation was a better predictor of learner’s understanding than the results of the product evaluation. In this paper as an additional analysis, we reported the analysis of relation between process evaluation and product evaluation. The difference between the evaluation methods could result in a case where there is zero score for process evaluation and full score for product evaluation. However, analysis results confirmed that there is a strong correlation between the two evaluation methods with $r$ value $= 0.8$. These results suggest that support for the process of submap re-composition is a promising way for learners to recompose a good concept map. We further explore the usability of the process evaluation to the teachers, and it shows a high rate of agreement of the usefulness and appropriateness of the method. Based on the results, we are developing a formative assessment and support environment for concept map re-composition.

Keywords: Concept map, concept map assessment, process evaluation, Kit-build

1. Introduction

Concept map, developed by Novak in 1972 (Novak & Musonda, 1991) is widely accepted as a promising activity for learning and evaluation of learner’s understanding and facilitate independent learning and thinking (Hu & Wu, 2012). Studies of concept map in the assessment across multiple disciplines have conducted (Gregoriades et al., 2009; Kinchin et al., 2000; Yin et al., 2005). Many studies describe the assessment criteria such as structural and relational scoring (McClure & Bell, 1990; Novak & Gowin, 1984). These assessment methods are accepted widely but purposed to assess the product of concept map, not the process of concept map building.

Research in concept map assessment mentioned that to realize sophistication and adaptation of the learning and evaluation, assessment of the process of concept map building
is an important issue. A study by Jablokow (2015) suggests the importance of collecting data about learner’s actual mapping process that could reveal an insight about cognitive aspect implicit in a concept map building that related to learner’s understandings. Concept map building process also shown that produced map quality has a relationship with the order of created propositions (Srivastava et al., 2021) and sequential mapping (Chiu & Lin, 2012). However, to realize the observation of the concept map building process is still a big issue, because most proposed method to evaluate the concept map quality and the process of concept map building relies upon human-dependent, qualitative approaches to conduct the observation (McAleese, 1998; Rautama et al., 1997; Wong & Lian, 1998) which generally lacks consistency and a demanding time and efforts.

In this research, we proposed an automatic process evaluation method and conducted an experimental evaluation of the method to evaluate its relationship with the quality of the produced map. The proposed method is intended for a re-composition type of concept maps, not for a general one where learner builds a concept map from scratch. We applied the proposed method in Kit-Build (KB), a concept mapping re-composition framework that allows knowledge sharing between teacher and learner and allows automatic assessment of learner’s concept map (Hirashima et al., 2011, 2015). It is known that KB automatic assessment could be used to assess learner’s map quality and the validity has been checked to be similar with other scoring methods (Wunnasri et al., 2018). As an assessment of learner’s understanding, KB map score has proven to have a high correlation with learner’s actual understanding (Yoshida et al., 2013).

Since the components of building a concept map are provided in the KB, it is possible to record the created propositions that are matched in comparison with the propositions from teacher’s build map. To realize the process evaluation, we use the concept of sub-map as a substructure in a concept map. We define several sub-maps with a set of semantics in the teacher’s map and define a good process by re-composing the sub-map as a series of activities. Then we analyze the data from learner’s KB concept map activity in respective to the defined sub-maps. We already confirmed that the process evaluation has a relation with test scores and therefore related to learner’s understanding (Rismanto et al., 2023).

In this paper, we report the results of the analysis of the relation between the process evaluation and learner’s map quality, that is, the product evaluation. If there is high correlation, it means that the quality of the map can be improved by supporting the process. To realize the analysis, the process evaluation results were compared with the product evaluation results from the data of a practical use of KB in an undergraduate class. We further explore the usability of the process score to the teacher, in the context of evaluating the process evaluation from the teacher’s perspective. To realize it, the process evaluation results were compared with the product evaluation results from the data of a practical use of KB in an undergraduate class. We further explore the usability of the process score to the teacher, in the context of evaluating the process evaluation from teacher’s perspective.

Therefore, the research questions for this study are as follows: Does the concept map building process have a relation to the product score (RQ1)? And does the process score is appropriate and useful from the perspective of teachers (RQ2)?

2. Background

2.1 Concept Map Assessment

Concept map is a widely recognized method to visualize knowledge in a graphical format (Novak & Gowin, 1984). It is a simple but powerful representation of knowledge to support the learning activity between teacher and learner. To represent a concept map, a graph is constructed by using only two symbols: Nodes that represent concepts, and links that represents relationship between concepts (Cañas et al., 2016; Dwi Prasetya et al., 2020). In an educational purpose, concept maps have been used such as for sharing meanings between people, planning the problem-solving process, representing, and assessing the learner’s knowledge structure.
Many methods and have been studied and proposed in utilizing concept map as an assessment tool (Plotnick, 1997; Yin et al., 2005; Zheng et al., 2019). A person who can well understand the meaning of words in the concept map, called “rater”, manually works to assess the concept map. Novak and Gowin (1984) proposed a structural scoring. This method gives a high score for the correctness of hierarchy level and each valid crosslink. It is because ordering the concepts into hierarchy and connecting the crosslinks can reflect learner’s creative thinking. Hence, this method is called structural scoring.

Other assessment methods that consider the meaning of a proposition instead of the structure have been studied. McClure and Bell (McClure & Bell, 1990) proposed the relational scoring. This method works by checking the relationship possibility between each proposition, concept label appropriateness and compatibility between labels. The reliability and validity of the method has also been studied. McClure et al. (Mcclure et al., 1999) investigated the concept map assessment by requesting students to construct a concept map using provided concepts and creating the link words to connect between concepts. Raters then scored each individual map separately. Result shows that the relational scoring method has a close relation with structural scoring method, using criteria map (teacher-build map).

2.2 Kit-build Framework and Automatic Assessment

Kit-build (KB) is a framework in concept map knowledge sharing and automatic assessment (Hirashima et al., 2011, 2015). In this framework, the learner creates a concept map by recomposing existing components called “kit” which consists of nodes and links. The nodes and links are components that decomposed from an “expert map” (teacher-build map). The teacher constructs the expert map by extracting from the learning material. This method of concept mapping is a re-composition concept map, which means that instead of creating their own concept map from scratch, learners recompose it from the kit. The map constructed by learner (learner map) then superimposed with the expert map to determine the matching propositions.

There are two main tasks in the KB framework (Yamasaki et al., 2010). The first is the segmentation task, in which the teacher is requested to prepare the teacher-build map, called the expert map. The expert map is extracted from the learning material and is the teacher’s expression of a comprehension of the learning topic. The expert map is then submitted to the system and decomposed into “kit” which contains a list of concepts and links from the expert map. The kit then provided to the learners to be recomposed into a learner map. Hence, when utilizing the kit, students are just required to identify the provided components and link them, no component creation is required. Figure 1 (left) shows an example of an expert map and a kit.

The second task is restructuring task. Learners are asked to restructure the concept map from the provided kit, creating the learner-build map, or learner map. Figure 2 (left) shows an example of learner map. After completing the restructuring, the learner map was then uploaded to the system for the analysis. The system then evaluates the learner map by superimposing it with the expert map. The automatic assessment works by matching each learner’s proposition with the expert map propositions.

![Figure 1. Teacher-build map, or expert map (left) and Kit (right).](image-url)
Propositions from the learner map that match with propositions from the expert map will be given a score. The total score is the evaluation result of the concept map product, called “map score”. Map score is the percentage of the matching propositions number from the learner map compared with the total amount of propositions from the expert map. Figure 2 (right) shows the learner map superimposed with the expert map. The solid lines are the correct propositions, and the dashed lines are the incorrect and lacked propositions in comparison with the expert map. Map score has been known with a high relationship with the learner’s understanding (Yoshida et al., 2013) which implies that learners with a high map score tend to have better knowledge compared with the ones with low map score. As an assessment of concept map, map score has been validated with other concept map product scoring method and comparable to the manual assessment methods (Wunnasri et al., 2018).

![Figure 2. Learner map (left) and learner map superimposed with expert map (right).](image)

2.3 Process Score

The process score proposed in this study aims to assess learner’s process in the KB concept map recomposing activity. The current scoring system of KB, the map score, is based on the produced concept map built by learners using the provided components (Jablokow et al., 2015). Our proposed method purposed to evaluate learner’s actual concept mapping process by analyzing the sequence of connected propositions in KB concept map recomposing activity. As mentioned by Chiu and Lin (2012) and Srivastava (2021), the sequence in concept map building process is related to the map quality built by learner where good process produces a good quality concept map.

In realizing the automatic process evaluation of concept map building, we analyzed it as a set of sequences, grouped by sub-maps. Sub-map is subset composed of closely related propositions in a concept map. Sub-map defined the concept map’s substructure (Gerstner & Bogner, 2009), that demonstrate the teacher’s deep understanding as the subject specialist (Kinchin & Alias, 2005; Roberts, 1999). Therefore, achieving it is an important target for learners. The re-composition process suggests that a learner has a solid comprehension of the substructure indicated by the sub-map when they continually recompose a collection of propositions in a sub-map.

The sub-maps are defined by the teacher after creating the expert map. Generally, a sub-map contains propositions that have a close meaning or extracted from the same subdomain of knowledge. As mentioned by Schneider et al. (2021), segmenting the concept map into a group of sub-map can be based on a particular knowledge domain of the concept map. This is in line with spatial continuity principle stated in Moreno and Mayer (1999) and Schroeder and Cenkci (2018), in which the information in a concept map is spatially ordered nearby within a learning material. Figure 3 (left) shows the example of sub-maps in a concept map.

After the sub-map is defined, learners are then given the task to recompose the kit back into the learner-build map, or learner map. In this stage, learners are not able to see the sub-map. The process scoring is transparent from the learners, as they will be performing the recomposing activity as usual. After completing the activity, learners will submit the map to the system and be given the map score and process score.
To calculate the process score, we analyze the sequence of connected propositions. We scanned the learner’s propositions that are sequentially connected as a set and corresponds to the propositions in each sub-map. The percentage of matched propositions sequence set for each sub-map is then calculated, which is, the number of connected propositions as a set divided by the number of propositions in a sub-map. As long as the propositions that are connected sequentially belong in a sub-map, the order of the proposition sequence in that sub-map is not considered.

For example, there are two sub-maps: Sub-map 1 (SM1) contains proposition 1 (P1), P2, P3. And SM2 contains P4, P5, P6. If a learner’s connected proposition sequence is P2, P3, P4, P5, P6, P1 then the calculation will be as follows:

P2, P3 connected sequentially as a set and belong to SM1. Therefore, the SM1 score is 2/3 * 100 = 66.67. Moving forward, P5, P4, P6 connected sequentially as a set and belong to SM2. Therefore, the SM2 score is 3/3 * 100 = 100. The process score is the average of the total score of each sub-map (SM). So, this learner will get a process score of (66.67 + 100) / 2 = 83.33. Figure 3 (right) shows the illustration of the process score calculation.

![Figure 3. Sub-map in a concept map (left) and example of process score calculation (right).](image)

3. Method

3.1 Experiment details

The study subjects were 28 second year undergraduate students majoring in Information Technology at State Polytechnic of Malang, Indonesia. Subjects have no prior experience or knowledge in building a concept map and in using the KB system. At the first session of the experiment, subjects were given a tutorial about creating a concept map and the usage of KB system. At the next session, the teacher explained about the learning materials in the form of lecture using power point slides. Further, subjects were tasked to perform KB concept map recomposing activity. After completing the recomposing, subjects upload the learner map into the system to be scored for a map score and process score. The total time for the experiment sessions is 60 minutes.

The learning subject is Polymorphism in Object Oriented Programming (OOP). The material was taken from the OOP course module released by the team teaching of the course at the Information Technology department at the State Polytechnic of Malang, Indonesia. The expert map was extracted from the module by the author of this paper and presented to the three teacher that has teaches OOP subjects for undergraduate students. The expert map was then discussed and modified according to the team teaching’s input, and the sub-map was also defined. The expert map consists of 10 propositions, 10 links, 11 concepts and five sub-maps.

3.2 Analysis
To assess the relation between KB concept map building process reflected by process score and learner’s map quality reflected by map score, a correlation test between process score and map score was performed. This assessment is based on the research by Chiu and Lin (2012) and Srivastava et al. (2021) that mentions the strong relationship between concept map building process and the quality of concept map product. Furthermore, a multiple linear regression analysis between map score and process score was conducted to check whether the process score has a strong influence and can be a predictor to the map score.

Detailed analysis of the sub-maps was carried out to confirm that the process analysis could provide useful information for teachers in recognizing parts of the concept map in relation to learner’s map quality. Based on the sub-structure of concept map that is defined by sub-maps (Gerstner & Bogner, 2009; Kinchin & Alias, 2005), we conduct this analysis to evaluate how each sub-map correlates to the quality of concept map product. The information can be used for teachers to provide adequate feedback for learners to improve their quality of produced concept map. To analyze the parts of the concept map that have a strong influence on learner’s map quality, correlation test was performed between each sub-map score and map score. Multiple linear regression analysis was performed to confirm that the highest correlated sub-map score with map score, can also be a predictor to the learner’s map quality.

To evaluate the teacher’s perspective on the reasonableness and perspective of the process scoring method, a questionnaire for teachers that have taught OOP subjects for undergraduate students and have familiarities with KB concept mapping has been conducted. The questionnaire was asked after observing the experiment and the method of process scoring and evaluating the results. We explained to the teachers how the process evaluation method works and its result, and the relation between process score and product score. Therefore, the teachers answered the questionnaire based on the relation between process scoring result and product scoring result. The first part of the questionnaire is about the reasonableness of the scoring method (Wunnasri et al., 2018). And the second part is their perspective of usability of the sub-map evaluation.

4. Result and Discussion

To check the relationship between process score and learner’s map quality which reflected by the map score, a Pearson correlation test was performed, and the result can be seen in Figure 4. According to the result of the correlation test, the R value of 0.8 means that process score has a high correlation with the map score. With p-value of 3.4e-07 which is less than significance level alpha = 0.05, it is concluded that process score and map score are significantly correlated.

This result means that learners with a high process score tend to have a better map quality compared to the ones that have a low process score. In a cognitive point of view, the process of recomposing the kit into concept map, reflected by the proposition sequence, could be an indication to assess the understanding level of learners, as mentioned by Yoshida et al. (2013) that the map score has a high correlation with learner’s understanding.

![Figure 4. Correlation between map score and process score](image-url)
Multiple linear regression analysis was carried out between map score and process score, to check that the concept map building activity process can be a predictor of learner’s map quality. Result shown in Table 3 concluded that with a p-value of 3.45e-07 which is less than significance level alpha = 0.05, process score has a strong influence on the map score. Therefore, the process score can be used to predict the map score. This result combined with the correlation test result concluded that the process score could represent the learner’s map quality reflected by the map score.

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 16.3931 | 10.3655 | 1.582 | 0.126 |
| sms         | 0.985    | 0.1454 | 6.774 | 3.45E-07 *** |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Detailed analysis on the sub-map has been carried out to check the significant sub-map. Pearson correlation was used to check the correlation between map score and sub-map 1 score (SM1), SM2, SM3, SM4 and SM5. The results shown in Figure 5. It is shown that sub-map 2 (SM2) has the highest correlation with map score compared with the other sub-maps with R value = 0.63. Multiple linear regression analysis between map score and each sub-map has been carried out to check the significance of the SM2 on the influence with the map score. Result shown in Table 3 concluded that the SM2 has high influence with the map score compared to the other sub-maps and can be a predictor of the map score.

The correlation analysis result combined with the multiple linear regression analysis result shows that the SM2 is the significant sub-map, that is part of concept map that has a strong influence on learner’s map quality. Learners that can connect propositions in the SM2 as a set of sequence tend to have a good map score, therefore, have a good understanding of the learning subjects. In the context of formative assessment, teachers can use this information to provide adequate feedback emphasized on the knowledge domain represented in the SM2 to further improve learner’s map quality. Improved map quality leads to a better understanding of the learning subjects, since it represents the learner’s knowledge (Novak & Gowin, 1984).
Furthermore, a questionnaire to the three teachers that have taught OOP subjects for undergraduate students and have familiarities in KB concept mapping was given to understand their perspectives on the KB process scoring. This questionnaire was asked after observing the experiment, the method and process evaluation result and sub-map analysis result. Thus, the teacher answered the questionnaire based on the relationship between process scoring and product scoring results. The appropriateness of process score evaluation method was asked in the first part, and the usability of sub-maps evaluation was asked in the second part.

The result of the teacher’s perspective questionnaire shown in Table 4. In the first question about the appropriateness of the process scoring method, it is noted that the process score is an appropriate method to evaluate learner’s process in KB concept map re-composing activity. In the second question, as an alternative to the map score, one teacher disagrees with the process score as an alternative assessment for KB. When asked for the reason, the teacher suggested that the process score is to be used as a complementary score to the map score. This is so that the results of the process score become an additional consideration for a teacher in assessing learner’s understanding. In the third question about the detailed analysis of the sub-maps, the teachers noted that the evaluation of the sub-maps in relation to product score is an appropriate method to investigate parts of sub-maps that highly correlated with product score. Finally in the last question, it is noted the provided information can be used for the teacher to provide adequate feedback in emphasizing the important domain knowledge represented by the highest correlated sub-map.

Based on the results of the questionnaire, we concluded that the teachers agreed with the necessity of process evaluation. The teachers agreed that the submap based process evaluation method is a reasonable one. Based on these agreements, they judged the process evaluation scores would be useful for learning support. Therefore, based on these results, we concluded that the process evaluation is a promising method, and it is worth continuous investigation.

Table 3. Multiple linear regression analysis between map score and each sub-map

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|---------|
| (Intercept) | 16.74635 | 12.23106   | 1.369   | 0.1848  |
| sm1      | 0.22916  | 0.12185    | 1.881   | 0.0733  |
| sm2      | 0.26448  | 0.11127    | 2.377   | 0.0266  *|
| sm3      | 0.05553  | 0.14394    | 0.386   | 0.7034  |
| sm4      | 0.21471  | 0.11667    | 1.84    | 0.0793  |
| sm5      | 0.17877  | 0.09257    | 1.931   | 0.0664  *|

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4. Questionnaire on the teacher’s perspective

<table>
<thead>
<tr>
<th>Questions</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Appropriateness of process score evaluation method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process evaluation on kit-build is an appropriate method to evaluate learner's process in kit-build concept mapping re-composing activity.</td>
<td>66.70%</td>
<td></td>
<td></td>
<td></td>
<td>ME - 33.30%</td>
</tr>
<tr>
<td>Process scores can be used as an alternative assessment for kit-builds other than map scores.</td>
<td></td>
<td>33.30%</td>
<td></td>
<td></td>
<td>66.70%</td>
</tr>
<tr>
<td><strong>Usability of sub-maps evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis of sub-maps is an appropriate method to find out which parts of a concept map affect the map-score.</td>
<td></td>
<td></td>
<td></td>
<td>66.70%</td>
<td>33.30%</td>
</tr>
<tr>
<td>By understanding the important sub-maps and sub-domain of knowledge, teacher can provide appropriate feedback to improve student understanding.</td>
<td>33.30%</td>
<td></td>
<td></td>
<td></td>
<td>66.70%</td>
</tr>
</tbody>
</table>
5. Conclusion

Assessing the learner’s process in the KB concept map re-composition activity reveals that process has a relationship with learner’s map quality. In this study, we proposed a concrete technique to assess the process of concept map re-composition in KB concept mapping. Based on the KB map score that can represent learner’s map quality, correlation analysis between process score and map score shows that process score has a high correlation with map score. Therefore, it answered the first research question (RQ1) that process in KB concept map building activity has a relation with learner’s map quality.

Detailed analysis on the sub-map shows that there is a significant sub-map that has a high correlation with the map score. This analysis could provide information for teachers to recognize part of the concept map that represents a knowledge domain that has the importance of the learner’s map quality. Teachers can use this information to provide adequate feedback to improve learner’s map quality and therefore, improve their understanding. The questionnaire result shows that the process scoring method is adequate and provides useful information in the perspective or teachers. Thus, it answered the second research question (RQ2).

Based on the questionnaire result, we concluded that the process evaluation is a promising method, and it is worth continuous investigation. The limitation of the proposed method is the order of sub-map is not considered. Therefore, further study to conduct more analysis and experiment could expand another aspect of KB concept map re-composing process. As a future works, the bigger size of concept map and number of participants can strongly confirm the conclusion of this study. Furthermore, building an environment to support concept map re-composition based on adequate process is one potential study to improve leaner’s map quality.

Acknowledgements

We would like to thank all the people who prepared and revised previous versions of this document and Indonesia Endowment Funds for Education (LPDP) for supporting this study.

References


On-demand lectures with humor and questions using avatars
Satoshi Toyota\textsuperscript{a} & Asuka Terai\textsuperscript{**}
\textsuperscript{a} Future University Hakodate, Japan
\textsuperscript{*}g2122042@fun.ac.jp
\textsuperscript{**}aterai@fun.ac.jp

Abstract: On-demand lectures are becoming increasingly popular due to COVID-19. Meanwhile, the numbers of students (viewers) who cannot concentrate on on-demand lectures are also increasing. One reason for this is boredom. Humor and questions are added to the lecture scenario in order to resist boredom during viewing on-demand lecture videos. Thus, a method was proposed to generate humor with “Boke (jokes),” “Tsukkomi (corrections),” and questions for important words in lecture scenarios and create lecture videos for demonstration by two avatars. Furthermore, to verify the validity of the proposed method, two experiments were conducted with 304 and 562 participants. The results of the experiments suggest that including humor and questions in on-demand lectures, demonstrated by two avatars, had the effect of inhibiting boredom.

Keywords: Humor, Questions, Lecture scenario, Avatar

1. Introduction

In recent years, the need for on-demand lectures has increased due to COVID-19, and on-demand videos have also been utilized in flip-style lectures after COVID-19. However, the need to devise a way to prepare for lectures has arisen because some students find it difficult to concentrate on lectures attended from home. According to Osugi (2021), on-demand lectures can guarantee knowledge gain; however, it is difficult to implement independent, interactive, and in-depth learning, and it remains difficult to change the way a person perceives things and thinks. Furthermore, Matsushima (2020) pointed out that on-demand lectures require significant time for teachers to prepare, which indicates that the burden on the faculty may increase in preparing for on-demand lectures.

In a previous study on substitute lectures by robots (Hiyori et al., 2018), a system in which a robot gave lectures at a university on behalf of lecturers was proposed. However, there are still some problems faced by lectures given by robots, such as mechanical nature and boredom. Meanwhile, according to Watanabe et al. (2020), it has been suggested that an avatar of the learner’s own liking may be able to generate interest in the content and motivate the learner to take the next class.

In addition, a previous research (Yamashiro et al. 2020) on incorporating humor into educational settings has shown that adding toilet humor to Kanji drills can change students’ grades and concentration. According to Yamashiro et al., students who used question books that incorporated humor showed an increased percentage of correct answers compared to students who used traditional question books. In addition, when concentration was measured by fixed-point camera imaging, it was found that students who used humor-infused materials concentrated on the paper from an earlier stage. Furthermore, Suzuki et al. (2021) proposed a comic script composition method for presentation talks. In this previous study, the authors found that incorporating a comic dialogue-like script structure into a presentation improved audience awareness, empathy, and familiarity. Furthermore, the effect of questions on viewers in lecture videos has been a focus of research. In one previous study (Yashio, & Araki, 2019), the insertion of questions for the viewer in a lecture scenario was proposed to generate lecture videos that do not bore the viewer.

Previous studies suggest humor and interaction during viewing on-demand lectures may be effective to resist boredom. However, adding humor or questions as interaction to the
regular lecture scripts increase the burden on faculty in preparing for on-demand lectures. This paper proposes a method to generate humor and questions for important words in lecture scenarios and create lecture videos for demonstration by two avatars to improve the problem of “boredom with on-demand lectures.” A lecture scenario with humor is generated by adding “Boke (jokes)” and “Tsukkomi (corrections)” about important words to the lecture scenario. The lecture scenarios follow a conversational format in which two avatars proceed with humor to soften the strict atmosphere. In addition, questions regarding important words are added to the scenarios for more active learning. The avatar asks students questions, giving them time to think, thus creating an active learning environment that makes the lecture less boring.

2. Method

We propose a humor method consisting of “Boke,” “Tsukkomi,” and questions regarding important words in the lecture scenarios that are input by a lecturer as a user. Two avatars representing teacher and student characters (“teacher avatar” and “student avatar”) demonstrate these scenarios so that students do not get bored with the lecture videos. The teacher avatar reads the lecture scenario aloud. The student avatar makes a joke as a “Boke,” and immediately after the “Boke,” the teacher avatar makes a correction of the joke as a “Tsukkomi.” In addition, the student avatar asks the student (viewer) questions regarding important words in the lecture scenario. Furthermore, the student avatar makes “Aizuchi (reaction)” along with the teacher avatar’s talk to make their dialogue more natural.

The lecture avatars are placed at both ends of the lecture video screen, as shown in Figure 1. Vroid Studio was used to generate the two avatars, with the teacher on the right side and student on the left side (see Figure 1). In addition, the lecture scenario is read aloud using Coe Font by avatars with an avatar lip-sync (the avatar's mouth moving in time with the voice) using 3tene. In addition, Open Broadcaster Software (OBS) was used to overlay a lecture slide and avatars to create a single lecture video, and subtitles were added using Adobe Premiere Pro (Figure1) because the text reading using synthetic voice Coe Font, may be difficult to hear the text.

![Figure 1. Lecture video by two avatars](image)

2.1 Generation of “Aizuchi (reaction)”

Templates (“Ho-ho,” “So-Nanda,” “Naruhodo,” and “Hee” (Japanese “Aizuchi” expressing an agreement like “I agree”) were created, and one of them was randomly selected to make each “Aizuchi.” According to the previous research (Kamiya et al., 2010), after auxiliary verbs (“Desu,” “Masu,” etc.) there is a 27.5% probability that “Aizuchi” is given in a dialogue. Thus, assuming that the part of speech at the end of a sentence in a lecture scenario is almost always an auxiliary verb, an “Aizuchi” was inserted after the end of the sentence in the lecture avatar's talk at 27.5% probability.

2.2 Generation of “Humor”
The proposed method aims to make the lecture scenario humorous based on the Japanese comedy style “Manzai,” in which two comedians have a humorous conversation. In “Manzai,” one speaker says a “Boke (joke)” and the other speaker corrects the joke by “Tsukkomi.” Thus, a “Boke (joke)” and “Tsukkomi” were generated for the important words in the lecture scenarios.

2.2.1 Generation of “Boke”

In a previous study on the automatic generation of comic dialogue, Manzai (Yoshida, & Hagiwara, 2012) proposed three types of “Boke (joke)” : “Kotoba-Asobi-Boke (word game),” “Kajo Boke (excessive joke),” and “Mujun Boke (conflicting joke).” In this study, we adopted “Kotoba-Asobi-Boke (word game),” which can easily be applied to important words in a lecture.

In “Kotoba-Asobi-Boke (word game),” the Japanese target word is romanized and one consonant is replaced with another, as shown in Figure 2. The generated “Boke” must be a real word that is commonly heard in daily life, or else the meaning of the “Boke” will be unknown. Therefore, a textbook vocabulary corpus (Kyokasho Corpus Goiyou) was used. The textbook corpus contains words used in elementary, junior high, and high school textbooks, and each word is written in Japanese in Kanji and Romaji alongside its frequency of use. The consonants of the target word are randomly converted and compared with the words in the textbook corpus, and those that match the reading of the word in the textbook corpus with the highest frequency of use are output as “Boke.” Words such as “death” that do not easily lead to laughter were removed from the “Boke” candidates. Words that consist of multiple morphemes (e.g., “Shusoku teki shikoh (thinking of convergence)” ) are split into their individual morphemes by MeCab(Kudo et al., 2004), one morpheme (e.g., “Shusoku (convergence)” ) other than a suffix (e.g., “teki (of)” ) becomes the target, and the consonant of the target morpheme is changed randomly to create the “Boke (e.g., “Shutoku teki shikoh (thinking of learning).”

![Figure 2. Generation method of “Boke”](image)

2.2.2 Generation of “Tsukkomi”

As a “Tsukkomi” against the generated “Boke,” we adopted the “Setsumei Tsukkomi (explanation correction)” used by Yoshida and Hagiwara (2012). The “Setsumei Tsukkomi” provides negations and explanations for the “Boke.” To generate explanations for “Boke,” the meaning of the concept word used as the “Boke” is searched in the Japanese WordNet. Japanese WordNet is a Japanese dictionary in which the relationships between concepts and definitions of words corresponding to concepts are registered. For words that consist of multiple morphemes, only words that are generated as a “Boke” for target morphemes are searched. For example, when “Shutoku (learning)” is generated as a “Boke” for “Shusoku (convergence)” in “Shusoku teki shikoh (thinking of convergence),” “Shutoku (learning)” is searched for in the Japanese WordNet to generate an explanation.

“Setsumei Tsukkomi” consisting of negations and explanations is generated by randomly combining three types of negations (“Why are you (boke)?”, “No (boke),” and “(not saying anything)”) and three types of explanations (“(boke) is (explanation),” “That is (explanation),” and “That is (explanation), isn’t it?”) (e.g. “No Shutoku (learning). Shutoku (learning) is acquisition of knowledge or ability.”). These are based on the canned sentences in Yoshida et al.’s [8] and were modified as “Tsukkomi” appropriate for the teacher character because the avatar playing the role of a teacher performs the “Tsukkomi.”
2.3 Question Generation

The questions are generated based on a previous study (Yashio, & Araki, 2019). Yoshio and Araki extracted questions have nouns that similar to the important words in the lecture from the Corpus of Japanese classroom Lecture speech Contents (CJLC) (Tsuchiya et al., 2008) and generated questions by replacing the nouns in the questions with the important words in the lectures (Figure 3). Instead of CJLC, this study used the Nagoya University Conversation Corpus (Fujimura et al., 2012) and Corpus of Japanese Teacher Speech (Nihongo Kyoshi Hatsuwa Corpus), which is a corpus of transcribed conversations, mainly by graduate students, that could be used to generate more comprehensible questions for students. The Corpus of Japanese Teacher Speech contains questions actually asked in lectures, and thus, it is more realistic as questions that appear in lectures. Questions that have similar nouns to the important words in the lecture were selected in these two corpora and replaced the nouns that appear in the questions with the important words that appear in the lecture. For example, for the important word “Shutoku teki shikoh”, a question"Shutoku teki shikoh suru koto arimasuka? (Do you ever do thinking of convergence?) " was generated.

3. Evaluation Experiments

Two experiments were conducted to validate the proposed method.

3.1 Experiment1

The purpose of the evaluation experiment was to verify the effect of humor consisting of “Boke” and “Tsukkomi” for “comprehensibility” and “boredom resistance.”

3.1.1 Method

A total of 304 people, who were paid cloud workers recruited by Lancers, participated in the experiment as a web-based survey (Males=187, Females=111, No Answer=6, Average age=42.29, SD of age=9.26). Three types of lecture videos were used: (1) a lecture video in which a teacher avatar reads a lecture scenario without a student avatar (Teacher), (2) a lecture video in which a teacher avatar reads a lecture scenario and a student avatar only gives an “Aizuchi” (Teacher+Student), and (3) a lecture video in which a teacher avatar reads a lecture scenario and “Tsukkomi,” and a student avatar gives an “Aizuchi” and “Boke” (Teacher+Student with Humor). In this study, the subjects were not shown the original lectures because it is difficult to generally compare between lectures given by avatars and by the actual teachers. Three types of existing scenarios and slides (Lectures 1, 2, and 3) that were actually used in a cognitive science lecture related to creativity at Future University Hakodate were used. At the end of the questionnaire, asked to rate their learning experience of cognitive science on a seven-point scale. The mean of the learning experience of cognitive science was 1.36, and the standard deviation was 0.89. The lecturer, who generated the scenarios and
slides and was one of the authors, selected two important words for each scenario. To avoid an order effect, the participants were divided into three groups (Group 1: 93, Group 2: 105, and Group 3: 106), and they watched the videos in the order of Lectures 1, 2, and 3 under different conditions depending on the group (Table 1). The length of each lecture video is 1:26 to 1:53 and viewers had time to answer questions after each video.

<table>
<thead>
<tr>
<th>Lecture Type</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1</td>
<td>Teacher</td>
<td>Teacher+Student</td>
<td>Teacher+Student with Humor</td>
</tr>
<tr>
<td>Lecture 2</td>
<td>Teacher+Student</td>
<td>Teacher+Student with Humor</td>
<td>Teacher</td>
</tr>
<tr>
<td>Lecture 3</td>
<td>Teacher+Student with Humor</td>
<td>Teacher</td>
<td>Teacher+Student</td>
</tr>
</tbody>
</table>

Immediately after watching each video, they were asked to rate how easy they found the videos to understand (comprehensibility) and how resistant to boredom they were (boredom resistance) on a 7-point scale using GoogleForm. After watching all videos, the participants were asked to respond to four-choice confirmation questions regarding the two important words in each video. One of the important words was a target word for “Boke” under the Teacher+Student with Humor condition.

3.1.2 Results

Using R, a two-way ANOVA (group × condition) was conducted to assess comprehensibility. An interaction was observed at the 1% level (F(4,602)=41.5, p<.01), and multiple comparisons using Holm’s method revealed significant differences (Figure 4(a)). The results show significantly lower average ratings for the videos last viewed, regardless of condition (5%). Similarly, two-way ANOVA was conducted for boredom resistance. An interaction was observed at 1% (F(4,602)=9.92, p<.01), and multiple comparisons using the Holm method revealed significant differences (Figure 4(b)). The average rating was higher for the first video viewed than for the last in the Teacher condition and the average rating was significantly lower for the last video viewed than the others in Teacher+Student condition (5%). However, there is no simple main effect was observed in the Teacher+Student with Humor condition.

![Figure 4. Average ratings of (a) comprehensibility and (b) boredom resistance](image)

To compare correct response rates for questions about the important word with “Boke” and about the other word without “Boke,” a McNemar test was conducted using the results of the confirmation test under Teacher+Student with Humor condition. The number of
participants for the results of the questions are shown in Table 2. The rates of correct responses to questions regarding important words with and without “Boke” were 68.1% and 61.2%, respectively, which were significantly different at the 1.0% level ($\chi^2(1)=7.23$, $p<.01$).

<table>
<thead>
<tr>
<th>Without “Boke”</th>
<th>With “Boke”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>41</td>
</tr>
</tbody>
</table>

3.1.3 Discussion

The comprehensibility results show higher average ratings for the videos viewed in earlier order, regardless of condition. Similarly, the boredom resistance results indicate the effect of viewing order. The average rating was higher for the first video viewed in each group regardless of the condition, and the rating was significantly lower for the last video viewed in each group, regardless of the condition. Meanwhile, simple main effects were observed among groups in the Teacher and Teacher+Student conditions, but not in the Teacher+Student with Humor condition. These results suggest that the inclusion of humor in a lecture video makes it difficult to be influenced by the viewing order and may make the participants concentrate on the lecture until the end.

The results of the confirmation test under Teacher+Student with Humor condition show that the rate of correct responses for the important words with “Boke” was significantly higher than without “Boke.” The inclusion of “Boke” for the important words may have an effect on the recognition of the words.

3.2 Experiment 2

The purpose of the evaluation experiment was to verify the effect of questions for “comprehensibility” and “boredom resistance.”

3.2.1 Method

A total of 641 people, who were paid cloud workers recruited by Lancers, participated in the experiment as a web-based survey, of which 79 participants also attended Experient 1 and the remaining 562 participants were valid (Male=308, Female=253, No Answer=1, Average Age=42.25, Age SD=10.72). The survey was conducted from 17:52 Oct. 21 2022 to 14:58 Oct. 23 2022. Same as Experiment 1, the target lecture was a cognitive science course on creativity. The mean of the learning experience of cognitive science was 1.39, and the standard deviation was 0.95. Four types of lecture videos were used: (1) a teacher avatar reads a lecture scenario without a student avatar (Teacher), (2) a teacher avatar reads a lecture scenario and a student avatar asks a question about an important word and gives an “Aizuchi” (Teacher+Student with Q), (3) a teacher avatar reads a lecture scenario and a student avatar asks two questions about important words and gives an “Aizuchi” (Teacher+Student with 2Qs), and (4) a teacher avatar reads a lecture scenario and “Tsukkomi,” and a student avatar gives an “Aizuchi,” “Boke,” and question (Teacher+Student with Q+Humor). Three types of existing scenarios and slides were used (Lectures 1, 2, and 3), as in Experiment 1. Same as experiment 1, the subjects were not shown the original lectures.

To avoid order effect, the participants were divided into six groups (Group 4: 95, Group 5: 97, Group 6: 93, Group 7: 96, Group 8: 89, and Group 9: 92), and they watched the videos in the order of Lecture 1, 2, and 3 under different conditions depending on the group (Table 3). It is assumed that questions are selected by the lecturer (user); thus, each
question used in Experiment 2 was selected from 10 generated candidate questions according to the results of a pilot experiment in which seven undergraduate and graduate students were asked to rate the questions on a 5-point scale. As in Experiment 1, the length of each videos were 1:28 to 1:54, and viewers had time to answer the questions after each video.

<table>
<thead>
<tr>
<th>Lecture Type</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Group 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1</td>
<td>Teacher</td>
<td>Teacher</td>
<td>Teacher+Student with Q</td>
</tr>
<tr>
<td>Lecture 2</td>
<td>Teacher+Student with Q</td>
<td>Teacher+Student with 2Qs</td>
<td>Teacher+Student with Q+Humor</td>
</tr>
<tr>
<td>Lecture 3</td>
<td>Teacher+Student with Q+Humor</td>
<td>Teacher+Student with Q+Humor</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

“Q” is one question added to the scenario and “2Qs” is two questions added.

3.2.2 Results

Using statistical software R, a two-way ANOVA (group x condition) was conducted for comprehensibility using groups with the Teacher+Student + Q condition (Groups 4, 6, and 8). An interaction was observed at the 1% level (F(4,548)=12.3, p<.01), and multiple comparisons using Holm’s method revealed significant differences (Figure 5(a)). In addition, two-way ANOVA was performed using groups under the Teacher+Student + 2Qs condition (Groups 5, 7, and 9). An interaction was observed at the 5% level (F(4,564)=25.1, p<.01), and multiple comparisons revealed significant differences (Figure 5(b)). Both of the results show lower average ratings for the videos last viewed in Teacher condition (5%).

![Figure 5](image)

Figure 5. Average ratings of comprehensibility of (a) groups with Teacher+Student + Q condition and (b) groups with Teacher+Student + 2Qs

( ** p<.01, * p<.05, The numbers below the bars indicate the order of viewing )

Two-way ANOVA (group x condition) was conducted for boredom resistance using groups with the Teacher+Student + Q condition (Group 4, 6, and 8). An interaction was observed at the 1% level (F(4,548)=5.88, p<.01), and multiple comparisons using Holm’s
method revealed significant differences at the 5% level (Figure 6(a)). In addition, two-way ANOVA was performed using groups under the Teacher+Student + Qs condition (Group 5, 7, and 9). An interaction was observed at the 1% level ($F(4,564)=25.1$, $p<.01$), and multiple comparisons revealed significant differences (Figure 6(b)). Both of the results show show significantly lower average ratings for the videos last viewed in Teacher condition (5%).

Figure 6. Average ratings of boredom resonance of (a) groups with Teacher+Student + Q condition and (b) groups with Teacher+Student + 2Qs

( ** p<.01, * p<.05, The numbers below the bars indicate the order of viewing )

To compare correct response rates for questions about the important words with and without a question, McNemar’s test was conducted using the results of the confirmation test under the Teacher+Student with Q condition (Groups 4, 6, and 8). The number of participants for the results of the questions is shown in Table 4. The rates of correct responses to the questions about important words with and without question were 64.6 and 52.9 respectively, which were significantly different at the 1.0% level ($\chi^2(1)=16.33$, $p<.01$).

Table 4. Results of confirmation test about important words with and without a question under Teacher+Student with Q condition

<table>
<thead>
<tr>
<th>Without question</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>129</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>81</td>
</tr>
</tbody>
</table>

To compare correct response rates for questions about the important words with a question and with a “Boke,” a $\chi^2$ test was conducted using results of the confirmation test under the Teacher+Student with Q+Humor condition (Groups 1, 2, and 3) and Teacher+Student with Q condition (Groups 4, 6, and 8). No significant differences were observed at the 5.0% level.

Furthermore, to compare correct response rates for questions about the important words with a question and with a “Boke” in the same video, McNemar’s test was conducted using results of the confirmation test under the Teacher+Student with Q+Humor condition. The number of participants for the results of the questions is shown in Table 5. The rates of correct responses to the questions about important words with a question and with a “Boke” were 54.4 and 60.0%, respectively. There was a significant difference at the 1.0% level ($\chi^2(1)=7.45$, $p<.01$).

To compare effects of questions and “Boke” when the other question was included in the same lecture video, a $\chi^2$ test was conducted using results of the confirmation test under the Teacher+Student with Q+Humor condition and Teacher+Student with 2Qs condition (Groups 5, 7, and 9). The rates of correct responses to the questions about important words with a question and with a “Boke” were 60.6 and 67.9%, respectively (Table 6), and there is significant trend between them ($\chi^2(1)=2.84$, $p<.1$).
Table 5. Results of confirmation test for important words with a question and with a “Boke” in the same lecture video under Teacher+Student with Q+Humor condition

<table>
<thead>
<tr>
<th>With question</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>257</td>
<td>49</td>
</tr>
<tr>
<td>Incorrect</td>
<td>80</td>
<td>176</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Answer</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Boke”</td>
<td>168</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>89</td>
</tr>
</tbody>
</table>

3.2.3 Discussion

The comprehensibility results show higher average ratings for the videos viewed first and lower average ratings for those viewed last in the viewing order, regardless of the condition. A simple main effect between groups in each condition was observed under the Teacher and Teacher+Student with Q+Humor conditions but not under the Teacher+Student with Q or Teacher+Student with 2Qs condition.

Similarly, the results for boredom resistance indicate an influence of viewing order. However, the results of Group 8 (Teacher+Student with Q+Humor, Teacher, and Teacher+Student with Q, in that order) do not show an order effect. Furthermore, a simple main effect between groups in each condition was observed under the Teacher condition but not under the other conditions.

This suggests that the inclusion of a question text directed to the participants regarding important words in the lecture video scenario may make it difficult to be influenced by the order in which they watch the video and may allow the participants to concentrate on the lecture until the end.

The results of the confirmation test for the important words with and without a question show that the rate of correct responses for the important words with questions was significantly higher than without questions. The inclusion of questions for the important words may affect word recognition.

Furthermore, the results of comparison between correct response rates for questions about the important words with a question and with a “Boke” in the same video show that the rate of correct responses for the important words with question was significantly higher than with “Boke.” This suggests that the inclusion of questions for the important words may have more of an effect for the recognition of words than the inclusion of a “Boke.” However, the results of comparison between effects of questions and “Boke” when the other question was included in the same lecture video show that a “Boke” is more effective than a question.

4. Conclusion

This paper proposed a method to generate humor consisting of “Boke,” “Tsukkomi,” and questions for important words in lecture scenarios, demonstrated by two avatars, to improve the problem of “boredom with on-demand lectures.” Two experiments were conducted to verify the validity of the proposed method. It was found that the addition of either a “Boke” or question made the participants less bored and also showed that the inclusion of a “Boke” and question for important words may have an effect on the recognition of words. However, since the content of the videos used in the experiments was only cognitive science, it is necessary to examine whether the proposed method is also effective for learning other content, e.g., mathematics. Furthermore, since the lecture video in this experiment lasted for
shorter than two minutes, it is unclear whether it will be an on-demand lecture that can be practically. It is necessary to conduct experiments using lecture videos that are similar in length to the on-demand lectures that are actually used. In addition, the evaluation of boredom resistance was subjective, and an objective evaluation is not yet possible. It is necessary to verify the use of physiological indices, such as the number of blinks. In addition, because including many sets of “Boke” and “Tsukkomi” increases the time of the lecture video and is considered impractical, we did not set a double humor condition in which two sets of “Boke” and “Tsukkomi” were included in the experiments. Hence, it is necessary to conduct additional experiments with a double humor condition to compare between effects of question and “Boke” when another “Boke” is included in the same lecture video.

References

CoeFont, https://CoeFont.cloud, (LastView:2023/5/1)
3tene Studio, https://3tene.com/studio/, (LastView:2023/5/1)
Kyokasho Corpus Goihyo, https://clrd.ninjal.ac.jp/bccwj/freq-list.html, (LastView:2023/5/1)
Adapting Noticing Framework to Analyze Learner's Reasoning in VR-simulated complex scenarios

Herold PC\textsuperscript{a*}, Chandan DASGUPTA\textsuperscript{b}

\textsuperscript{a}Educational Technology IDP, IIT Bombay, India
\textsuperscript{b}Educational Technology IDP, IIT Bombay, India
*herold@iitb.ac.in

Abstract: Car drivers learning to make decisions in Indian on-road scenarios is challenging. We designed and simulated a complex on-road scenario using virtual reality technology to help learners experience alternate perspectives of their driving actions. Our research focused on the cognitive influences on decision-making and adapted the noticing framework as the analytical lens. Four participants drove a car on virtual Indian city roads accompanied by a virtual conversational agent to elicit their reasoning while making decisions, followed by a stimulated recall protocol. We recorded each participant's actions and timestamped conversations with the agent and researcher. We inductively coded the noticing episodes and interpreted six types of conceptual relationships between the designed elements in the VR scenario. These interpreted codes could help us analyze the changes in a learner's noticing patterns after taking alternate perspectives. This paper highlights the usefulness of the adapted noticing framework for future studies with multiple complex VR scenarios.

Keywords: VR simulation, Noticing framework, On-road scenarios, Learner's reasoning analysis

1. Introduction

A decision made by a human in complex scenarios requiring a time-bound response involves reasoning processes as characterized by naturalistic decision-making models (G. A. Klein, 1993). A vital component of that process involves the learner noticing and assessing the critical cues within the complex situation. In our research, we contextualized a complex scenario based on the frequently occurring events on Indian public city roads, simulated in a Virtual Reality environment. Existing work on VR-based decision-making and situation assessment has majorly modeled extreme accident situations as complex scenarios (Benvegnù et al., 2021) rather than frequently occurring events. VR interventions with affordance to take alternate perspective-views of an event (Kang et al., 2012) for decision-making have used ethical/moral frameworks (Sütfeld et al., 2017) to analyze their actions. These analytical lenses include both cognitive and affective influences on decision-making. Since capturing the affective responses of learners while driving is challenging, we focused our research on cognitive analytical methods. This scoping led us to the noticing framework in the education research domain. The existing noticing frameworks are predominantly derived in the teacher education and learning sciences field within a classroom setup (Amador et al., 2021). In this paper, we chose the noticing framework (Lobato et al., 2012) from learning sciences due to the possibility of mapping the conceptual element in our VR scenarios with the adapted components of the framework. In the VR intervention we developed, "PERSPECS," the learners/participants drive a VR-simulated car on Indian public city roads, making quick decisions at critical junctures, followed by taking an alternate perspective view of their actions. Our broader research objective is to investigate the effects of virtual reality perspective-taking on the learner's noticing behaviors in complex scenarios while driving a car on Indian roads. But, in this paper, we investigated a scoped-down research question: "How
does the adapted noticing framework help capture the learners' assessment of critical cues within the complex VR scenarios?".

2. Noticing framework

Research on Noticing as a construct is derived from Goodwin's work on professional vision (Goodwin, 1994). For our study, we are considering the definition from the learning sciences group (Lobato et al., 2012), which says - Noticing refers to selecting, interpreting, and working with particular features or regularities when multiple sources of information compete for students' attention. While assessing a complex situation on Indian roads, noticing the critical cues from various sources of information across the sequence of events is essential for decision-making (G. Klein, 2008). Lobato's noticing framework maps well with our context, which consists of the following components: (i) Conceptual Relationships, (ii) Perceptual Features, (iii) Focusing interactions, (iv) Task features, and (v) The Nature of the Activity. We adapted the framework by modifying the domain-specific definitions for each component detailed in the following sub-section.

2.1 Framework adaptation

In our context, a 'Conceptual Relationship' is a relationship between different elements in a complex scenario that can be abstracted to signify a concept. The 'Perceptual Features' in our context help drivers orient toward the conceptual elements in the situation and notice the conceptual relationships between them. 'Focusing Interactions' refers to the virtual conversational agent (co-passenger) interactions with the participants, detailed in the next section, which can emphasize the various conceptual relationships. The 'Task Features' are the multiple on-road complex scenarios designed. The 'Nature of Activity' refers to the participatory roles of the driver and other actors in the VR environment, which regulates their possible actions. Particularly, the 'Conceptual Relationships" forms the basis of analyzing the noticing episodes coded from the participant's conversations with the agent, detailed in the discussion section.

3. Design of VR intervention

The noticing framework and the user study interviews conducted by us with ten car drivers commuting in Indian cities guide the design of the VR-simulated complex scenarios. The interview method was semi-structured, with questions to recall and reflect on their driving experience, for example – “Can you share any instance where you had to break a traffic rule but felt your decision was right?” The VR-simulated car also includes the design of a conversational agent guided by the scaffolding framework of pedagogical agents (Azevedo et al., 2004). We developed the scenario with the "Unity-3D" software. The scenario is experienced in VR using an "HTC-VIVE" head-mounted display and controlled using a physical steering wheel and brake/accelerator, as shown in Figure 1.

3.1 VR-simulated Complex scenario

A complex scenario is a scoped version of complex problems defined by (Funke, 2010), characterized by Complexity, Connectivity, Dynamics, and In-transparency in his work on cognition. In our research, we scope our scenarios that imbibe the characteristics of complex problems. In such complex scenarios, the decision-maker needs to notice and assess many elements involved in events that are changing dynamically and respond within 3-4 seconds. This response action results in immediate consequences on fellow human beings in that public scenario. The elements of each complex scenario are primarily characterized into the following four types – (i) Action Target - The action target of a scenario is humans who could bear the consequences of a participant's driving actions, (ii) Action Rule - The action rules of a scenario communicate the road rules to be followed by the driver in that context. (iii) Action Persuader
The vehicles in the participant's vicinity which persuade or pressure the participants to make a quick decision, and (iv) Action Choice Primer - The vehicles in the participant's vicinity that could prime them on the available choices to act in that context. We derived these characteristics for each design element in the VR scenario from the affinity-map inferencing done on user study interviews.

A brief to the "Old man scenario" shown in Figure 1 – "You are on the way back to your home in your car to attend a dinner you had planned, and you have stopped in front of a traffic signal as the signal is RED (Action Rule). A few pedestrians and an old man (Action Target) are crossing the road. As the old man is halfway across the road nearing your car, the signal turns GREEN, the vehicles by your side start moving (Action Choice Primer), and vehicles behind you start honking (Action Persuader)." In this public scenario, the complexity is in assessing the situation with multiple events unfolding. The response time to decide and act in this complex scenario is around 3-4 secs. We validated this scenario using a survey questionnaire with 30 Indian drivers, who rated multiple Indian on-road scenarios based on the following factors - familiarity with them and level of difficulty in deciding an action.

3.2 Conversational agent

The Wizard-of-Oz method (Dahlbäck et al., 1993) is adopted to deliver the appropriate dialogues and animation gestures of the co-passenger to the participant, where a researcher will use the desktop keyboard to trigger them, as shown in Figure 1. This method allows the researcher to trigger multiple dialogues like - "Hey, back there, how did you decide what to do?" based on the response from the participant.

4. Method

The primary research question we are analyzing in this study is – "How does the adapted noticing framework help capture the learners' assessment of critical cues within the complex VR scenarios?". The participant's conversation with the conversational agent during the VR intervention and the participant's conversation with the researcher during the stimulated recall (Lyle, 2003) are coded as per the adapted noticing framework to interpret patterns for the primary research question. We used qualitative analysis with an inductive coding method (Cohen et al., 2017) to derive inferences.

4.1 Participants and procedure

The participants for the study were adults in the age range of 21 - 29 years with a minimum car driving experience of one year on Indian roads. Four participants volunteered for the study and signed the necessary consent forms approved by the institutional review board on campus. None of the participants had prior experience in using a VR car simulation. Participants take 3-level training inside the PERSPECS VR environment for 15 minutes to get accustomed to the car controls, gain familiarity with the city buildings and roads, and get familiar with the co-passenger avatar and self-avatar within the VR environment. After training,
the participant experiences the complex scenario, i.e., the "old man scenario," as per the following sequence of steps for another 15 minutes. Firstly, as detailed in the design section, the participant drives the car, encounters the traffic lights with the old man and other pedestrians crossing the road, and acts at the critical decision-making point. After the action, the co-passenger has a conversation right after the critical decision-making point. Following the conversation, the participant experiences their action replay from the old man's perspective inside the VR environment (taking a perspective view is part of the broader research objective and not detailed in this paper), followed by a conversation with the co-passenger again. Ending the VR experience, the participant removes the VR HMD. The researcher then replays the screen recording on the desktop and conducts a stimulated recall protocol to reiterate their reasoning at the critical points in the scenario for 7-10 minutes.

4.2 Data sources and analysis

The data sources include (i) The video recording of the participant experiencing the VR intervention, (ii) The audio recording of the participant conversing with the co-passenger and the researcher, and (iii) The screen recording of the entire VR action-play by the participant. We used a qualitative analysis with an inductive coding method to derive inferences for our research question- "How does the adapted noticing framework help capture the learners' assessment of critical cues within the complex VR scenarios?". The transcripts between the co-passenger, participant, and researcher are encoded where the participant notices a conceptual relationship (CR) between the elements within the VR scenario. For instance, the conversation between the co-passenger and the participant in Table 1 is a noticing episode, which we then coded inductively into their various types.

Table 1. A conversation excerpt between the co-passenger and a participant

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-passenger</td>
<td>What all thoughts were there in your mind before acting back there?</td>
</tr>
<tr>
<td>Participant-3</td>
<td>I thought he is old; he was quite slow, so I thought I should give him time to cross the road.</td>
</tr>
</tbody>
</table>

In this conversation, the participant noticed the conceptual relationship of walking pace between the old man and the other pedestrians crossing the road. Noticing a conceptual relationship becomes one of the reasons for the participant's decision to give time for the old man to cross the road. Hence, similar conversations are "noticing" episodes that form the analysis unit. Following this, we inductively interpreted the types of noticing episodes and categorized them as shown in Table 2. We refined our categories through repeated discussions, where we examined their descriptions.

5. Results and Discussion

We encoded the conversations between the four participants, co-passenger, and researcher resulting in 29 noticing episodes. Out of these, ten noticing episodes were after the critical decision point, at an average of 2.5 episodes per participant, and 19 noticing episodes were after perspective-taking, amounting to an average of 4.75 episodes per participant. We interpreted these noticing episodes and categorized them into six types based on the conceptual relationships noticed between the designed elements of the VR scenario. The following subsection details the categorized noticing episodes with examples of such reasonings spoken out by the four participants.

5.1 Noticing episodes with conceptual relationship

The six types of noticing episodes were interpreted and categorized based on the conceptual relationships between the designed elements in the VR scenario are summarized in Table 2.
The top two types of noticing episodes observed across the four participants are "You Driving (YD) + Action Rule (AR)" with nine episodes and "Action Persuader (AP) + Action Rule (AR)" with six episodes. The following dialogue is an instance of the noticing episode interpreted as the type "Action Persuader + Action Rule" from the reasoning conversations of participant 4 – "Ya, like this, even in the real world it happens, and I usually stop my vehicle, and if the vehicle behind don't notice I sometimes roll the window and signal using my hand that 'wait' so they usually understand. or sometimes I signal from the rearview mirror the vehicle behind can see my eyes when I give an expression, they usually understand and wait that something is there in front and they avoid honking. But if it was a teenager, they honk again and again."

Table 2. Types of noticing episodes interpreted with description.

<table>
<thead>
<tr>
<th>Noticing types</th>
<th>Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action Target (AT) + You Driving (YD)</td>
<td>When the participant reflects and talks on conceptual relationships like the old man's (AT) response in relationship with their vehicle's (YD) behaviors towards him and its immediate consequence towards the old man.</td>
</tr>
<tr>
<td>Action Persuader (AP) + Action Target (AT) /Action Rule (AR)</td>
<td>When the participant reflects and talks on conceptual relationships like the behavior of the car behind (honking) (AP) in relation to old man's behavior (AT) or traffic signal (AR).</td>
</tr>
<tr>
<td>Action Choice Primer (ACP) + AT/AR</td>
<td>When the participant reflects and talks about conceptual relationships, like other vehicle's response (ACP) in relation to old man's (AT) behavior or traffic signal (AR).</td>
</tr>
<tr>
<td>YD + AP/ACP/AR</td>
<td>When the participant reflects and talks on conceptual relationships between their car (YD) and behaviors of the honking vehicle behind (AP) or other vehicles by the side (ACP) or change in traffic signal light (AR).</td>
</tr>
<tr>
<td>Retro (R) + AT/AP/ACP</td>
<td>When the participant reflects and talks on conceptual relationships between an object in the scenario and retrospective accounts (R) of that object from previous experiences.</td>
</tr>
<tr>
<td>AT + AR</td>
<td>When the participant reflects and talks on conceptual relationships like old man's (AT) behavior in relationship with other pedestrians while crossing the road (like walking pace) or in relation to changes in traffic signal light (AR).</td>
</tr>
</tbody>
</table>

The types of noticing episodes inductively interpreted and categorized in Table 2 based on the conceptual relationships will help critically analyze the participant's reasoning while making decisions in complex scenarios. For our broader research objective of investigating the effects of virtual reality perspective-taking, this approach will allow us to explore the change in noticing patterns across multiple VR scenarios. For instance, in the "Old man scenario," the noticing episodes of type "Action Target + You Driving" were interpreted differently by the participant before and after perspective-taking. One of the dialogues of Participant 4 before perspective-taking is – "Now what do I say about that, signal was RED, but then the old man was there, so I just stopped, and also there is a rule that you have to give first priority to the pedestrian. I am in a car so I can go easily but they are walking so". On the contrary, the dialogue of Participant 4 after perspective-taking is – "Actually in that view I was just judging myself, so at that time the car shook a bit, because at that time when I was moving right it was going too much right and likewise for left. So if I were an old man I would never cross the road seeing someone shake the car like this, thinking that they might hit me. Also,
the car came back and forth and I felt from the old man’s view that if it was the old man, then he would have got scared.”. These coded noticing episodes with six categories indicates that the adapted noticing framework used to design the VR intervention is meaningful and valuable for more extensive studies with multiple complex VR scenario in our context. The number of participants is a limitation of the study, and if the number is higher, there is a possibility of interpreting more types of noticing episodes in the future. One of the challenges in adapting the noticing framework is to identify the appropriate conceptual elements within our context and map them to the definition of noticing framework components.

6. Conclusion

Using advanced virtual reality technology, we developed complex scenarios that can show alternate perspectives to help car drivers learn decision-making in complex on-road scenarios. This paper discusses how we adapted the noticing framework, developed a complex VR scenario, and inductively interpreted the participant's reasoning by categorizing the noticing episodes. Four participants drove a car on virtual Indian city roads accompanied by a virtual conversational agent to elicit their reasoning while making decisions. We interpreted the noticing episodes into six types of conceptual relationships between the designed elements in the VR scenario. In future work, we can use the identified noticing characteristics to identify and analyze the patterns in noticing behaviors before and after taking perspective views across multiple complex scenarios in the Indian road context. This paper demonstrates the usefulness of the adapted noticing framework for such contexts and can be used by Learning Sciences and Education researchers for VR-based interventions.

References


Analysis of algorithmic strategy development in the development of computational thinking of upper elementary school students

Xiaowen WANG, Pinqi HU & Guang CHEN*
Faculty of Education, Beijing Normal University, Beijing, China
*guang@bun.edu.cn

Abstract: Computational thinking has become a thinking ability that we need to master in our daily lives. Programming is a way to cultivate computational thinking. This study focuses on the algorithmic links of senior primary school students in programming activities, puts forward a hierarchical table of algorithmic strategies, and analyzes their strategic development paths according to their programming behaviors. In this paper, Swift Playgrounds, developed by Apple Company, is selected as the tool of programming activities, and three levels are selected as the teaching content in the three theme activities of Learning Programming 1 in Swift Playgrounds software. The target audience of the activities is the senior (fourth to sixth grade) students in primary schools. This paper studies the design of the micro-generation method, carries out programming activities based on computational thinking for students through pre-test, practice, post-test and migration stages, and collects data by interview and classroom observation so as to analyze the development path of students' algorithm strategies in different stages and the change track of algorithm strategies in each stage. This study provides some support for teachers' teaching in programming teaching activities.

Keywords: Programming teaching activities, computational thinking, micro-genetic method, algorithm strategy

1. Introduction

In the context of the rapid development of information technology, society has put forward new requirements for people's abilities and accomplishments. People should learn to understand some problems encountered in social production, life and learning from the perspective of computer science, and be able to use the basic principles of computer science and advanced technical tools to solve complex and difficult problems (Tedre & Denning, 2016).

Computational thinking is a key research topic in computer education. Different researchers have different definitions of computational thinking, but the definition of computational thinking proposed by Jeannette, which includes five elements: algorithms, decomposition, abstraction, generalization and debugging, has been widely recognized by researchers all over the world. Jeannette believes that computational thinking includes thinking activities such as problem-solving, designing systems, and understanding human behavior. This is not only a skill that scientists need to possess, but also a skill that every one of us needs to possess (Jeannette, 2006).

As one of the ways to cultivate computational thinking, programming education can cultivate students' systematic problem-solving and strategic thinking abilities. In the process of programming, students can not only learn some basic knowledge but also learn logical methods and problem-solving strategies through specific problem situations, so it becomes a powerful starting point for cultivating computational thinking (Winslow et al., 2018).
According to the cognitive development level and thinking development process of primary school students, primary school students above grade three are gradually transitioning from image generalization level to abstract generalization level, and their reasoning ability is also in a stage of rapid development. They can independently and logically demonstrate more complicated judgments and processes. Therefore, at this stage, students' computational thinking can be cultivated through programming teaching. For example, Saez-Lopez et al. (2016) through on-the-spot evaluation of the effect of using Scratch's visual programming language in classroom practice, explore whether students have made significant progress in learning programming concepts, logic and computational practice; Zhao & Shute (2019) asked students to play "Penguin Go" video game in the classroom, that is, to create a program to guide penguins to cultivate students' cognition and attitude towards the development of computational thinking skills through a regional approach; On the other hand, Chen-Chung et al. (2011) designed a simulated game to let students learn and use programming knowledge to solve situational problems, and explored the influence of simulated games on learning experience and problem-solving strategies related to programming knowledge. Pellas & Peroutseas (2016) showed students their participation in a 3D multi-user game environment by combining Second Life with Scratch4SL's 2D programming environment and used them to develop computational thinking skills in collaborative problem-based programming tasks.

To sum up, most of the current research on computational thinking focuses on the development and effect research of computational thinking activity courses, and the research on students' behavior in the process of computational thinking is relatively lacking. This study focuses on the cultivation of students' algorithm strategies in computational thinking activities. Starting with the basic programming knowledge, it studies the strategies and behaviors used by students when using programming knowledge to solve problems, and provides support for front-line teachers' programming teaching process.

Therefore, this study intends to solve the following problems:

(1) What are the algorithm strategies that students mainly use at different stages of different programming activities?

(2) What are the development paths of students' algorithms and strategies at different stages of the same programming activity?

2. Methods

In this study, the micro-generation method is used to study the changing process of students' algorithm strategies in the programming process, which is an exploratory preliminary study. This study mainly analyzes the development of algorithm strategy among senior primary school students in Swift Playgrounds-based programming activities, and further analyzes the path of students' strategy development according to the algorithm strategy grade table.

The whole teaching process of programming activities is divided into three thematic activities: command, function and For loop. Each thematic activity is divided into four stages: pre-test, exercise and post-test. After the completion of the three thematic activities, a complete programming project is carried out and used as a test in the migration stage. And through interviews and classroom observation to collect data. The interview method is mainly used to interview students after each stage in the process of teaching by the micro-generation method so that students can orally express what strategies are used to complete this part of the challenge and why such strategies are used to solve the problem. Classroom observation is mainly used to track and record the development of students' algorithm strategies in class so as to further analyze students' algorithm strategies.

The research object of this study is the students in senior grades (grade four to grade six) in primary schools, and they are recruited through online registration. A total of 15 students were recruited, including 12 boys and 3 girls, aged between 10 and 12, with an average age of 10.9 years and a standard deviation of 0.59. There are 12 fourth-grade students, 2 fifth-grade students and 1 sixth-grade student. During the implementation of the activity, a total of 15 students participated, but 2 students failed to complete all the courses. After the
implementation of the activity, a total of 15 students' learning data were received, of which 13 were valid. Among the valid data collected, from the point of view of school distribution, 9 people are studying in primary schools in Beijing, while 4 people finish their usual studies at home. In terms of gender, there are 10 boys and 3 girls, with an average age of 10.2 years and a standard deviation of 0.38.

This experiment is divided into three courses, and the content taught in each course is exactly the same. During the class, each student is recorded on an iPod Touch. At the same time, the iPad used by the students turns on the screen recording function, and a corresponding number of graduate teaching assistants are equipped for stage interviews and behavior tracking.

In this experiment, there are 15 subjects, who have completed 12 levels of programming game challenges in 3 theme activities, and designed 19 interview questions in the pre-test, practice, post-test and migration stages. Based on this, the obtained data is encoded. Combined with the specific programming behavior of middle school students in this research course, the algorithm strategies used by students in this programming process are summarized and classified, and according to the path and order of children's cognitive development from low-level strategies to high-level strategies, three experts in the field of psychology are invited to divide students' algorithm strategies into four levels, namely, zero-level strategies, first-level strategies, second-level strategies and third-level strategies. According to the different development characteristics of each level strategy, combined with students' specific algorithm behavior, the different level strategy characteristics of algorithm strategy are defined, and according to the video analysis and interview answers of 13 subjects' programming activities in Study 2, three experts teach coders to encode independently, finally improve the algorithm strategy level, and analyze the reliability according to the consistency percentage formula proposed by Holsti in content analysis. The r (reliability) of the algorithm strategy rating table is 0.9, and the reliability of the evaluation records of three judges, A, B and C, is greater than 0.9, which shows that the evaluation result of the main judge can be used as the result of content analysis, which achieves the consistency of coders. The strategy level rating table is shown in the following table:

<table>
<thead>
<tr>
<th>Algorithm Strategy Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>In the process of the algorithm, we rely on our own subjective judgment and experience to complete the algorithm.</td>
</tr>
<tr>
<td>Level 1</td>
<td>On the basis of subjective judgment, simple methods are used to complete the level goal, and the simple methods are limited to the commands involved in the level goal.</td>
</tr>
<tr>
<td>Level 2</td>
<td>On the basis of simple methods, use appropriate methods to complete the level goal, but it is limited to simpler scene goals.</td>
</tr>
<tr>
<td>Level 3</td>
<td>On the basis of appropriate methods, draw inferences from others and think independently about the best method needed to complete the checkpoint goal.</td>
</tr>
</tbody>
</table>

3. Results and Discussion

This chapter will analyze the data collected in the previous section. The study mainly uses qualitative analysis to investigate the changes in students' strategies and behaviors in the process of programming. It mainly uses encoding the video of students' entry into the classroom, and analyzing the interview contents in the pre-test, practice, post-test and migration stages to discuss the changes in students' strategies in the programming process. According to the micro-generation method, the breadth analysis of the same programming activity in pre-test, practice, post-test and migration stages is carried out. Mainly through the three theme activities, students' individual development routes are carried out, and the individual development roadmap is drawn according to the development of children's strategic development.
level at different stages. Based on this road map, this paper analyzes the migration degree of students’ algorithm strategies of three theme activities in the migration stage.

3.1 The development route of individual algorithm strategy in command theme activities.

From the roadmap of individual development of the algorithm, it can be seen that five people in the algorithm strategy of command theme activity have developed from Level 0 to Level 2; One person developed from Level 0 to Level 2 and then dropped to Level 0 and then to Level 2; Two people developed from Level 0 to Level 2 and then fell to Level 1 and then to Level 2; Five people directly developed from Level 1 to Level 2. There are four strategic development paths in total.

From the development path of algorithm strategy, all students use low-level strategy in the pre-test stage of command theme activities. In the practice stage, students can easily master this method after teachers teach the method of counting grids and planning the shortest route, so the second-level strategy is used the most frequently in the post-test stage, but a small number of students still use low-level strategy. In the migration stage, students’ algorithms all develop into high-level strategies, which also shows that in the algorithm strategy of command sentences, low-level strategy is a monotonically decreasing development model, and high-level strategy.

![Figure 1. Roadmap of individual development of command theme activity algorithm strategy](image)

3.2 The development route of individual algorithm strategy in function theme activities.

From the roadmap of individual development of the algorithm, it can be seen that four people in the algorithm strategy of function theme activity have developed from Level 1 to Level 3; Five people have developed from Level 1 to Level 3 and then dropped to Level 1; Four people developed from Level 1 to Level 2 and then fell to Level 1. There are three strategic development paths in total.

In the development path of algorithm strategy, all students use low-level strategy in the pre-test stage. Through the function method teaching in the practice stage, students' algorithm strategy has reached the second and third levels, but in the migration stage, Level 2 has dropped to Level 0, the most frequently used strategy level is the first-level strategy, and the third-level strategy is also used by students. It shows that in this part of the function, students do not fully grasp the definition and usage of the function and learn to draw inferences from others. The usage times of low-level strategies and high-level strategies are evenly distributed, and the students' algorithm strategies are not developing from low-level strategies to high-level strategies, but the development modes of low-level strategies and high-level strategies fluctuate and alternate constantly.
3.3 The development route of the overall algorithm strategy for loop theme activity.

From the roadmap of individual development of the algorithm, it can be seen that one person in the algorithm strategy of “For loop” theme activity has developed from Level 1 to Level 2; There are 12 people who have always maintained the level of secondary strategy. There are two strategic development paths in total.

In the theme activity of “For loop”, students’ algorithm and strategy have maintained a high level from the pre-test stage to the migration stage, even though the most frequently used strategy in the post-test stage is the first-level strategy. This is because in the post-test stage, it is necessary to customize the function and use the loop to write the code. According to the development map of algorithm and strategy in the function part, students can master the definition and usage of For loop skillfully in this part. So on the whole, the development of advanced strategies is constant.

Figure 3. Roadmap of Individual Development of Algorithm Strategy for Cyclic Theme Activity

Synthesizing the development paths of algorithm strategies of the three theme activities, we can get the following results: the proportion of algorithm strategies used by middle school students in command activities and “For loop” activities is getting higher and higher, and the overall trend is from low-level strategies to high-level strategies, but in the later period, low-level strategies have not all disappeared, and some students still use low-level strategies to solve problems; The strategy development path of the function part shows this point more clearly. The algorithmic strategy of the function part develops alternately. Even if students learn more advanced function knowledge in the practice stage, more students still choose lower-level codes to replace the function strategy in the migration stage, which shows that students’ cognitive strategies do not completely develop from low-level strategies to advanced strategies, but students subconsciously choose their most skilled and direct strategies by relying on their own knowledge and pre-concept experience in different situations.

4. Conclusion

This study uses the micro-generation method to study the changes of students' strategic behavior in programming activities, and makes a preliminary exploration in programming education. Through this study, it was found that students use different algorithm strategies when completing different programming activities. In such simple programming activities as command statements and For loop, students can use advanced strategies taught by teachers to complete tasks, which shows that students' algorithm strategies are constantly developing from low level to high level, and the migration effect of algorithm strategies is obvious. However, in such difficult programming activities as function statements, the algorithm strategies used by students before and after teacher intervention are not the methods taught by teachers, but their own subconscious strategies, which shows that the algorithm strategies have not developed from low level to high level in this link, and the transfer effect is not obvious. Therefore, the development of cognitive strategies is complicated, and students need certain cognitive resources for learning strategies, and choose the most appropriate cognitive strategies to solve problems according to specific situations.
The results of this study also show the differences in the development of students' cognitive strategies in mathematics. For example, in the "function" activity of this study, the teacher taught more advanced methods, but some students who had not been exposed to programming before did not master this method, so they still used their own methods in the subsequent task level, and in the cognitive development of arithmetic, all students can always complete the corresponding tasks according to the methods taught by the teacher. It shows that some students don’t have the pre-concept of programming activities, so they can’t finish coding and debugging according to the statement logic in programming in a short time. However, arithmetic activities are activities that students have been exposed to since childhood, and students can combine the relevant experience and pre-concept in the process of growth with the methods taught by teachers, and finally complete the corresponding tasks.

Therefore, on the whole, the development of children's cognitive strategies is complex and specific. Complexity is reflected in the fact that students do not completely abandon the low-level strategies formed by previous conceptual experience in the process of solving problems after learning more advanced strategies, but form a competitive relationship between low-level strategies and high-level strategies in the process of thinking. Although using more advanced strategies may increase the efficiency of solving problems, they are still in the process of accepting and understanding the more complicated and difficult advanced strategies they have just learned, which requires a lot of cognitive resources. It is impossible to complete a flexible application, so the specificity of cognitive strategy development is formed: students choose the most suitable strategy according to different environments to achieve the goal of solving problems.

There are also some shortcomings and defects in the design and implementation of this study. From the experimental link and process, this experiment focuses on the development process of students' algorithm strategy in programming activities, that is, based on whether they successfully complete the checkpoint task, while ignoring whether students have written wrong and redundant codes that will not affect the final task. Therefore, in the future research and development of this kind of programming software, the process can be designed more strictly, and if irrelevant code appears, it should also be prompted or challenged to fail, so as to ensure that students can successfully write and run the code not because of luck or subjective judgment. In addition, the selection of subjects in this experiment is not rigorous. Some students have already used Swift Playgrounds before taking part in the activity, so the results obtained in the pre-test stage and practice stage are not much different from those obtained in the post-test stage and migration stage, which leads to the unconvincing effect of the experiment.

References

393
Toward a Virtual Human Exhibit for Public AI Education

Ning WANG*, Timothy HURT*, Ari KRAKOWSKI*, Eric GREENWALD†, Omkar MASUR*, Boxi FU* & Chirag MERCENT*

*Institute for Creative Technologies, University of Southern California, USA
†Lawrence Hall of Science, University of California, Berkeley, USA

*nwang@ict.usc.edu

Abstract: It is important to prepare the world’s youth and the future workforce with fundamental knowledge of Artificial Intelligence (AI). Informal settings, such as museums, offer great opportunities in helping youth and the general public learn about AI. In this paper, we will discuss the design of a Virtual Human Exhibit that aims to communicate to the public about the capabilities and impact of AI through AI technologies used in virtual humans.

Keywords: K-12 AI education, learning in informal settings, virtual human

1. Introduction

Artificial Intelligence (AI) is a foundational technology that is transforming our work and daily lives. It is critical to prepare future generations with basic knowledge of AI. It is also critical for parents to learn about AI, as they are often the decision-makers for their children’s use of AI. Informal settings, such as museums, offer unique opportunities to reach out to two-generation learners – parent and child in learning about AI. In this paper, we discuss the design of a Virtual Human Exhibit that aims to communicate to the public about the capabilities and impact of AI through AI technologies used in virtual humans. The exhibit is design for science museums with visitors between 5- to 12-year-old, accompanied by their parents or guardians. The exhibit currently includes a number of activities that center around how virtual humans utilizes AI to realize functionalities similar to humans, such as recognize facial expressions. While the exhibit is still expanding through research and development, part of the exhibit has already been piloted at working museums with hundreds of visitors.

2. Virtual Human Exhibit

The Virtual Human Exhibit includes a number of interactive activities where the visitors can walk up, experience the AI-driven technologies used to help a virtual human character think, feel, speak, and act like a real human, and reflect upon the implications on their own lives. The current implementation of the exhibit includes 1-2-3 Smile! and Mirror Mirror, two activities where visitors interact with a couple of facial expression recognition AIs; Virtual Comedian, where the virtual human tells joke and use multimodal behavior analysis (e.g., from vision and speech) to infer if the visitors like the joke; Sensing Showcase, where virtual human converse with the visitors while the user interface displays visualizations of data processing behind-the-scenes (e.g., head pose and facial landmark tracking; speech recognition); and Silly Face, an activity discussed in detail in the following section.

Unlike learning taking place in formal settings, such as classrooms, a key challenge in designing educational experiences in informal settings, such as museum, is engagement. In this free-choice environments with many exhibits to choose from, visitors often spend little time deciding if they want to interact with an exhibit. This means that the design of the exhibit should (1) attract the visitors’ attention and pique their interest to interact with it. Given that the target learner population is 5- to 12-year-old children and their accompanying guardians, we designed the virtual human to be a character of similar age to the young visitors. Unlike
formal learning settings, with teachers to scaffold the learning and answer question, (2) the learning experience in a museum needs to be self-explainable, without the need for human assistance or intervention. Given that our visitors are primarily 5- to 12-year-old children, with limited reading ability, we designed the instructions used in the exhibit in simple languages. The instructions are given verbally instead of as text displayed on the screen. While the AIs driving the integrated virtual human are complex, explaining how such complex AI systems work is no small feat either, particularly to an audience of young age, without engineering background, in a short amount of time in a museum setting. We have developed a D-I-R (detect-interpret-respond) framework to illustrate the high-level process of how AI behind virtual humans work to serve as the underlying structure for scaffolding the understanding of such process (Greenwald, Krakowski, Hurt, & Wang, 2023).

One of the first exhibit activities that has gone through iterations of design, implementation, and user testing is *Silly Face*. In this activity, the visitors help the virtual human learn to recognize their silly and serious faces. The activity is organized in five phases:

- **In the initial **Introduction phase, the virtual human greets the visitors as they walk up to the exhibit (e.g., “Hi there! Would you like to try an A.I. activity with me? Press the red button to start!”) and briefly introduces the task (e.g., “For this activity, I am going to use A.I. to learn more about your facial expressions. Here, let me zoom in to your face.”). This phase clearly marks the beginning of the activity. In our earlier testing when this phase was not included, the young visitors were unclear about which phase the exhibit was in and what they should do to start it.

- **In the Data Collection** phase, the virtual human guides the visitors to pose silly and serious faces and captures images of those faces as training dataset (Figure 1, Top-Left). The virtual human provides the young visitors with clear instructions, examples, and encouragement to help them throughout this phase. The virtual human first provides an example silly face image: “*Do you know how to make a silly face? Here’s what my silly face looks like...*”, then asks the visitors to post silly faces: “*I would like to learn what your silly face looks like. Could you make some silly faces to help me learn?*” During the data capturing, the virtual human continues to speak to the visitors: “*Keep being silly; I learn better if I can see all of your different silly expressions.*” Such phrases serve not only as words of encouragements, but also as indications that the capturing phase is still ongoing. At the end of silly face images capturing, the virtual human concludes this round of capturing and introduces the next round: “Wow, I really like your silly faces! I think I have enough data to learn what your silly face looks like. I’d like to learn another one of your expressions. *Here is my serious face. Can you make serious faces?*”

- **In the Model Construction** phase, the virtual human learns from the images captured and builds a classifier to recognize the visitors’ silly and serious expressions. In addition to communicating verbally to the visitors, such as “Let me take a moment to learn the difference between your serious faces and silly faces.”, the user interface also plays animations where individual pictures are highlighted to one by one, to simulate that the virtual human is examining and learning from each image.

- **In the Model Testing** phase, the virtual human tries to recognize the visitors’ facial expressions using the learned model: “*Try making a bunch of different silly and serious faces to see if I am good at determining which one you are making.*” (Figure 1 Top-center and Top-Right). This phase lasts around 30 seconds.

- **In the final Reflection phase**, the virtual human reviews the images recognized as silly and serious faces with high confidence levels (Figure 1 Bottom-left and Bottom-center): “*I am pretty confident that you were being silly in this photo? Do you agree? Does this photo look similar to the silly faces you showed me earlier?*” and the ones with low confidence level (e.g., Figure 1 Bottom-Right): “*I had a hard time figuring out if you were making a serious face or a silly face in this picture? Why do you think I struggled?*”. The goal of the reviews is to help the visitors reflect how training data impacts the AI performances, and the limitation of the state-of-the-art AI (e.g., hard to identify facial expressions not seeing in the training dataset).
This activity captures the visitor inputs via a single physical button, which allows the visitors to start (or restart) the activity, and a webcam, which captures the visitors’ facial expressions. MediaPipe is used for face recognition tracking (Lugaresi et al., 2019). The facial expression classifier is implemented using MobileNetV3 (Koconce & Koconce, 2021). The virtual human’s speech is synthesized in real-time using Microsoft text-to-speech. The speech gesture and lip-synch are procedurally generated in real-time through NVBG (Lee and Marsella, 2006) and Cerebella (Lhommet et al., 2015) based on text of the speech.

3. Discussion
The Silly Face activity has recently studied through a 9-day exhibit at a science museum in the United States, with over 500 visitors. Observations from the project’s third-party independent evaluation team indicate that the exhibit created high levels of visitor engagement and dwell time, compared to other exhibits at the museum. Other activities from the exhibit, such as Mirror Mirror and Virtual Comedian have been piloted during science fairs at local high schools. Studies on these activities with target audiences at science museums are underway to gather data on age-appropriateness of these activities (e.g., can the young museum visitors understand the jokes told by the virtual human comedian?) and the guided interactions to help visitors grasp the AI behind virtual humans.

Acknowledgements
This research was supported by the United States National Science Foundation under Grant # 2116109. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References
Analyzing Learning Patterns and Potential Interventions in First-Year Compulsory Course at an Online University

Yasuhisa KATO

Professor, Tokyo Online University, Japan
*kato.yasuhisa@internet.ac.jp

Abstract: Online universities are becoming increasingly popular for their flexibility, especially for adult learners. However, the autonomy of online learning can lead to challenges in maintaining motivation and engagement. This study at Tokyo Online University examines the learning patterns of first-year students in a mandatory academic literacy course. It analyzes the relationship between students’ learning history and grades to identify patterns of success and failure in order to develop targeted intervention strategies. The literature review discusses MOOC attrition rates, self-regulated learning strategies, and the relationship between learning patterns and academic performance. The analysis clusters students based on engagement and performance, reveals distinct patterns between successful and unsuccessful students, and highlights early course engagement as critical. The study recommends introducing information on successful study strategies during new student orientation and sending targeted messages to disengaged students.

Keywords: Online universities, Learning patterns, Intervention strategies

1. Introduction

Online universities offer flexible, on-demand courses that are ideal for adult learners who want to retrain or upgrade their skills, freeing them from the constraints of time and place. As higher education increasingly moves online, it is important to understand the challenges and opportunities involved. A key challenge in online learning is the need for individual autonomy, self-regulation, and sustained motivation, without which learners may abandon their studies, a prominent issue in MOOCs (Hone & El Said, 2016).

Previous research highlights key factors for online student success, including time management and social presence (Kizilcec, Pérez-Sanagustín, & Maldonado, 2017; Richardson et al., 2017). This study aims to build on these findings to develop targeted intervention strategies for online students, inspired by Furukawa's (2022) analysis of MOOC student data. Despite advances in MOOC research, studies of learning behaviors in four-year online undergraduate programs are limited. Online students tend to be more intrinsically motivated and less socially connected than traditional college students (Kato, 2021; Kato, 2022b). Early engagement in courses correlates with higher final grades, as shown in a first-year course at an online university (Kato, 2022a). This paper focuses on analyzing learning patterns and exploring potential interventions for students, rather than detailing specific intervention implementations.

2. Literature Review

Studies of learning progress and attrition in online learning, particularly in MOOCs, are extensive. The average MOOC enrolls approximately 43,000 students and has a completion rate of 6.5% (Jordan 2014). Literature reviews indicate that the rigid structure of MOOCs may not be appropriate for students with diverse needs, and those who take courses selectively or slowly may be considered dropouts (Onah et al., 2014). Learning analytics
have been used to identify successful and unsuccessful learning patterns in blended environments (Tempelaar et al., 2015). A systematic review highlighted the importance of understanding different learning patterns in online higher education and suggested that factors that contribute to success in traditional settings may be less effective online (Broadbent and Poon, 2015). Kizilcec et al. (2017) explored the relationship between self-regulated learning strategies, learner behavior, and goal attainment in MOOCs. Furukawa (2022) analyzed MOOC programming learning logs and identified clusters of early learners to support based on characteristics such as number of video views and programming skills. Kato (2022a) found that early course completion has a significant impact on final grades, suggesting that early intervention for students who are falling behind could potentially reduce dropout rates.

3. Research Questions

The research questions are as follows:

- What learning patterns are associated with successful completion of a course during the first half of the quarter?
- What learning patterns are associated with failure during the first half of the quarter?

4. Analysis of Learning Patterns

Figure 1 displays a scatterplot comparing students’ first and eighth class completion dates with final scores, where a score of 60 or above, marked by a red line, signifies success. The graph indicates some early starters dropped out, while most who completed eighth class passed.

Figure 2 shows students clustered by first class completion date and final score: Group A started early, scored high, and did not need intervention. Group B started late, scored high, and benefited from the intervention. Group C started on time but failed and appeared to need time management intervention. Group D is a group that started slowly and lost credits, and how to intervene is an issue for the future.

Figures 3 and 4 show the full periodic learning histories for Groups A and C, who completed their first study within three weeks. The axes represent a quarter’s schedule and progress percentage, with 100% representing completion of all 8 classes. Due to similar trends, the analysis divides the students into two groups: those who passed (A and B) and those who failed (C and D). Clustering was done using the Python module "sklearn".

5. Future Discussion
In the future, the following information should be presented at new student orientation prior to the start of classes to help students understand their learning options and consider their own learning styles:

1. Take your first class as soon as possible
2. There are several ways to take classes. There are several ways to proceed with the class, either regularly or all at once in the first half of the class.
3. It is possible to catch up in the second half even if you fall behind in the first half.
4. If you don't start at the very beginning, you may not be able to finish.

Figure 3. Learning progress of Group A.  
Figure 4. Learning progress of Group C.

Acknowledgements

This research was supported by JSPS KAKENHI under Grant Numbers JP22K12303. I would like to thank the students at Tokyo Online University for their cooperation in this study.

References


A Comparative Study of Traditional and Augmented Reality-Based Engineering Drawing Instruction: Effects on Visualization Skills and Cognitive Load

Ajay Shankar TIWARI\textsuperscript{a}, Kaushal Kumar BHAGAT\textsuperscript{b*}

\textsuperscript{a,b}Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, West Bengal, 721302, India
\*kkntnu@hotmail.com

Abstract: Augmented reality (AR) is utilized in education to enhance students' skills and learning experience. We developed an AR-based application for engineering drawing (ED) courses, evaluating its impact on 3-D visualization skills and cognitive load. The study involved 392 first-year engineering students, with a control group using conventional methods and an experimental group using the AR application. Results showed that the experimental group outperformed the control group in 3-D visualization skills and experienced lower cognitive load. Students responded positively to the application, indicating that AR can enhance learning performance and experience in engineering education.

Keywords: Augmented reality, marker-based, cognitive load, system usability, engineering drawing

1. Introduction

Augmented reality (AR) technology enhances education by providing an interactive experience where computer-generated information augments the real world (Azuma, 1997). It helps students to visualize abstract concepts, increases motivation, and develops 3-D visualization skills. This study focuses on using AR in an Engineering Drawing (ED) course to improve spatial abilities and reduce cognitive load. A marker-based AR application was developed and evaluated using the System Usability Scale (SUS). Student feedback was collected to identify strengths and weaknesses of the application for future improvements. The specific research questions (RQ), this study intends to answer are as follows:

- RQ1: Can AR based learning improve students' visualization of 3-D objects in ED courses compared to the conventional instruction method?
- RQ2: Can AR-based learning in ED reduce students' cognitive load relative to conventional ED learning?
- RQ3: Is the augmented reality-based application easy to use for the ED course?
- RQ4: How satisfied are students with the AR-based learning method?

2. Related work

2.1 Augmented reality and its role in field of Engineering Drawing

AR integrates computerized information into the real world and has been implemented in various disciplines, including education, recreation, tourism, and healthcare. In education, AR has helped in developing immersive and interactive learning environments (Kaufmann & Schmalstieg, 2003). It enhances the classroom by providing real-time response, simulations, and visual aids, increasing student engagement and understanding. AR has improved learning outcomes and motivation across different educational contexts, making difficult ideas more accessible (Chang & Hwang, 2020). Interactive AR simulations has enabled students to
manipulate virtual objects and observe their behavior, simplifying the comprehension of abstract concepts (Bacca et al., 2014). ED is an important course in engineering education, but students struggle with visualizing 3-D objects from 2-D drawings (Pando Cerra et al., 2020). There have been studies where AR has enhanced learning by providing interactive experiences, generating 3-D models, offering visual aids, and enabling practical experience (Danakorn Nincarean et al., 2019; Martín-Gutiérrez et al., 2019). AR has also assisted in boosting motivation, engagement, and enjoyment in ED courses (Wei et al., 2011).

2.2 Cognitive load theory

Cognitive load theory (CLT) suggests that reducing the cognitive load enhances learning. Multimedia, like images and videos, can increase cognitive burden, but interactive media can facilitate learning (Sweller et al., 1998; Mayer & Moreno, 2003). AR decreases cognitive load by providing an immersive and interactive learning environment (Billinghurst & Duenser, 2012). Lee (2012) found that AR significantly reduced cognitive load in science learning compared to traditional methods. However, the effectiveness of AR in reducing cognitive load depends on system design and learners’ prior knowledge (Huang et al., 2020).

3. Methodology

3.1 Research design and sample

A pilot survey was conducted to identify challenging topics in ED. Participants included 53 undergraduates and 18 experts. The survey revealed difficult areas such as solids projection, cross-section, points, planes, and auxiliary planes. A true-experimental design was employed with 392 randomly selected students (291 males, 101 females) aged 19-22. The experimental group used an AR app, while the control group followed traditional methods. Each group consisted of 196 students. In this study, we developed a marker-based AR application using Unity\(^1\) software to enhance 3-D visualization skills in an ED course. The application included sections like orthographic view, projection of solids, and cross-section of solids. Users selected sections, pointed their device’s camera at markers, and viewed 3-D models. They could interact with the models by rotating, moving, and resizing them. The application provided various shapes and angles for visualization. Users were recommended to explore all sections within a week to fully utilize its features (see Figure 1).

\(^1\) https://unity.com/

Figure 1. User Interface of application.
3.2 Instruments

We administered a pre- and post-test containing 8 test items each to evaluate 3-D visualization skills, validated by subject expert (see Figure 2). Students were given 10 minutes to complete this test. Thereafter, we used a cognitive load questionnaire for analyzing students’ extraneous cognitive load and intrinsic cognitive load during the experiment (Hwang et al., 2013). SUS, created by Brooke (1996), was used for evaluating the usability of the AR application. It consisted of ten items on a five-point Likert-scale. Students were given 25 minutes to complete the post-test questionnaires. The Cronbach’s value for all questionnaires was greater than 0.7, which is acceptable (Barrett, 2001). We used a questionnaire to obtain qualitative feedback on our AR application. It consisted of 10 subjective questions through which students could discuss their experience using the application, thereby providing more support to our findings. The questions were like: “Do you have any previous experience of using AR app?”, “How much would you rate the app?”, “Did you find the AR useful?” etc.

Figure 2. Example of questions included in pre-test and post-test.

Figure 3. Experimental design
3.3 Procedure

Figure 3 depicts the experimental procedure of the investigation. The study consisted of three stages: (1) Both groups were obliged to complete the pre-test within the allotted time; (2) the experimental group was instructed to use the application to study ED for one week; and (3) both groups were required to complete the post-test. The experimental group was subsequently required to complete an SUS questionnaire and provide qualitative feedback.

4. Results

We examined the data using descriptive statistics. The learning outcomes of the experimental group and the control group were compared using t-test methodology. We used SPSS version-21 for result analysis.

An independent t-test was employed to analyze the results, and the effect size (Cohen's d for various sample sizes) was computed to demonstrate the extent of difference between the groups. In the pre-test, there was no significant difference between the control and experimental groups ($t = 0.30$, $p > 0.05$). In contrast, the post-test revealed a significant difference between the control and experimental groups ($t = 11.14$, $p < 0.05$). In addition, we conducted a t-test on the learning gains of both groups. There was a significant difference in performance between the control and experimental groups ($t = 6.64$, $p < 0.05$; see Table 1).

<table>
<thead>
<tr>
<th>Table 1: Means and standard deviations of the pre- and post-test for each group.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Pre</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Post</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Gain</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. * $p < .05$. SD = Standard deviation, MD = mean difference

An independent t-test was conducted to examine the impact of the AR-based learning approach on the intrinsic and extraneous cognitive load of students. The results showed a significant difference between the two groups in terms of mental load ($t = 16.78$, $p < 0.05$) and mental effort ($t = 15.86$, $p < 0.05$; see Table 2). These outcomes implied that the AR-based application had a significant impact on the students' intrinsic and extrinsic cognitive loads.

We calculated SUS score using technique proposed by Bangor et al. (2008). The total SUS score was 92.1237. Figure 4 depicts the SUS analysis and feedback obtained by students.

<table>
<thead>
<tr>
<th>Table 2: Means and standard deviations of cognitive levels of each group.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Mental load</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
5. Discussions

The study assessed the effect of AR technology on improving students’ 3-D visualization abilities in ED courses compared to conventional methods of studying ED. We created an AR application and encouraged the experimental group to study ED using the application while the control group studied ED using the traditional method. Four research questions were addressed in the study.

The first addressed the impact of AR-assisted ED learning on students’ 3-D visualization skills. The findings demonstrated that, as compared to traditional approaches, AR significantly improved 3-D visualization skill in the ED. This result was aligned with earlier research where AR was used to improve visualization skill of students from architecture background (Escudero et al., 2016). In addition, they were able to visualize points and lines in various planes, which was not possible during classroom instruction.

The effect of AR technology in reducing cognitive load in ED tasks is the focus of the second research question. The study found that students who exercised an AR-based application experienced substantially lower intrinsic and extraneous cognitive load compared to those who used traditional instruction methods. This result surpassed previous research that showed AR-based learning can minimize extraneous cognitive load in multimedia science (Lai et al., 2019).

The third research question addressed the usability of AR application for ED course. The average SUS score obtained revealed that our application is user-friendly and appropriate to ED courses. Previous research evaluating the utility of AR has shown equivalent results, indicating that AR-based applications can be employed in ED classrooms (Wijaya et al., 2019; Martin-Gonzale et al., 2010).

The final research question focuses on student satisfaction with the AR application. As indicated in the results section, students who had prior experience with AR used the application with ease, as did those who were experiencing AR for the first time. The application’s nature as a marker-based AR application allowed users to avoid the difficulties associated with marker-less AR applications (Wijaya et al., 2019).

6. Conclusions and future work

This study determined that AR-assisted learning in ED was able to enhance students’ 3-D visualization skills. The results also showed that students who used AR to study ED
experienced less cognitive load than those who used conventional methods. The application was subsequently evaluated using the SUS scale. It was also revealed that the application is user-friendly and that students can study ED using this application. In their response, students suggested that the software can be used as an alternative to Auto-CAD because it allows users to create custom designs. This turned out to be one of the limitations encountered by our application. In addition, the application was only compatible with handsets running the Android operating system and not iOS.

References


Wijaya, A. C., Munandar, M. W. A., & Utaminingrum, F. (2019, September). Usability testing of augmented reality for food advertisement based on mobile phone using system usability scale. In 2019 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 266-269). IEEE.
Measuring Self-regulated Learning Processes in Computer Science Education

Elizabeth B. CLOUDEa*, Ryan S. BAKERa & Maciej PANKIEWICZb

aPenn Center for Learning Analytics, University of Pennsylvania, USA
bInstitute of Information Technology, Warsaw University of Life Sciences, Poland
*ecloude@upenn.edu

Abstract: Self-regulated learning (SRL) is important for computer science education. Yet, students often do not have SRL skills to benefit their learning. In this study, we examined 187 (n=187) students’ SRL behaviors while they built programs with an automated feedback tool. Anchored in Winne and Hadwin’s (1998) COPES model of SRL, our results showed that novices used more operators to debug compiler errors, while more experienced programmers used more operators to debug non-compiler errors. Finally, a random forest classifier showed that prior knowledge was the most important COPES feature predicting learning gain, followed closely by the student’s perceived programming ability, use of evaluations with the automated feedback tool, and operators used to debug non-compiler errors on failed programs.

Keywords: Computer Science Education, Self-regulated Learning, Learning

1. Introduction

Computer science (CS) education requires students to develop self-regulated learning (SRL) skills (Prather et al., 2018); yet many students, particularly novices, do not have the SRL skills to benefit their performance (Arakawa et al., 2021; Loksa et al., 2016). SRL requires monitoring and adapting learning processes and strategies to pursue a goal. COPES (Winne & Hadwin, 1998), a widely adopted framework, emphasizes that feedback is the central mechanism driving SRL, adaption, and learning.

Many CS educators rely on automated assessment tools since they offer automatic feedback and evaluation (Chen et al., 2020), creating a feedback loop between the student and automated tool, possibly opening doors to support SRL. Arakawa et al. (2021) examined whether students struggled to initiate SRL as they programmed with an automated feedback tool. Results showed that often students ignored the feedback from the tool altogether, perhaps due to a lack of SRL skills (Arakawa et al., 2021). Loksa and others (2016) examined relations between coding errors and SRL strategies collected using a think-aloud protocol. While students programmed with an automated feedback tool, the results showed novices used SRL infrequently and often ineffectively for debugging errors. In contrast, experienced programmers used SRL during programming and they made fewer errors. Marwan et al. (2022) and Ko et al. (2019) found similar results to Loksa et al. (2016). Novice and experienced students differed in their use of SRL strategies and coding errors with an automated feedback tool. In sum, prior programming knowledge plays a role on how SRL manifests with automated feedback tools and its relation to code quality. More research is needed to explore relations between prior knowledge, SRL processes with an automated feedback tool, and learning outcomes in introductory programming.

2. Methods

Data were collected from 245 CS undergraduates (n=245) during a mandatory, first-semester programming course (C#) at a large University in Poland. Due to missing data for interest variables, students (n=58) were removed from our analysis, resulting in a subset of
187 students (n=187; 39% female). The course covered 146 tasks on 8 basic CS education topics. The automated assessment tool, runcodeapp.com (Pankiewicz, 2020) allowed students to submit code on a pre-defined set of test cases. Feedback was provided as a total score, compiler errors, and detailed results for each test case. To view detailed feedback for each test case, students had to click on a selected test case.

Participants completed a 5-point Likert scale (1=low, 5=high) to report their programming ability. An 8-item, multiple-choice pre-test was also administered to measure general programming knowledge (Mdn=38.75). The Likert scale and pre-test measures were used to assign students to a novice or more experienced group. A Wilcoxon test found pre-test scores were significantly higher for students who reported more programming experience (Mdn=65.88) than novices (Mdn=16.25), W=8079, p<.001. As such, students were assigned to the more experienced group (n=87) if they reported a higher programming ability (3-5), or to the novice group if low (1 or 2; n=100). A 9-item, multiple-choice test evaluating knowledge of the course's programming topics was administered at the middle (Mdn=55.56) and end of the semester (Mdn=66.67). Normalized learning gain (NLG) was used to calculate a normalized score of the max possible change from mid- to post-test (M=.09, SD=.38; see Marx & Cummings, 2007). To explore SRL with an automated feedback tool and its relation to learning, we grounded our work in Winne and Hadwin's (1998) COPES model (see Table 1). Log data on code and interactions with the automated feedback tool were collected.

Table 1. COPES construct descriptions and operational definitions.

<table>
<thead>
<tr>
<th>COPES</th>
<th>Description</th>
<th>Operational definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditions</td>
<td>Cognitive or external resources</td>
<td>• Self-reported programming ability.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• General programming knowledge scores.</td>
</tr>
<tr>
<td>Operators</td>
<td>Primitive information processing mechanics</td>
<td>The proportion of consecutive test cases for code that scored &lt;100: contained either a) compiler or b) non-compiler errors.</td>
</tr>
<tr>
<td>Products</td>
<td>Knowledge shaped by operators</td>
<td>Not analyzed in current study.</td>
</tr>
<tr>
<td>Evaluations</td>
<td>Feedback on discrepancies between products &amp; standards</td>
<td>Average clicks on results of resubmitted code from the automated feedback tool.</td>
</tr>
<tr>
<td>Standards</td>
<td>The criteria of success</td>
<td>Criteria for perfect program solutions.</td>
</tr>
</tbody>
</table>

3. Results

Do novices and more experienced programmers differ in the evaluations and operators used to debug failed programs with compiler and non-compiler errors?

The Benjamini and Hochberg method controlled for type I errors. A Mann-Whitney test found more experienced (Mdn=.5) and novice programmers (Mdn=.63) marginally differed in how many operators they used to debug compiler errors, W= 3687, adj. α=.05, p=.07122. The second test found that more experience programmers used more operators to debug programs with non-compiler errors (Mdn=7.29) than novices (Mdn=6.96), W=5220.5, adj. α=.025, p=.01845. A separate Mann-Whitney found more experienced (Mdn=1.03) and novice programmers (Mdn=1.04; p=.3852) did not differ in the use of evaluations on failed programs with the automated feedback tool (p>.05).

Which COPES features are important for classifying if students learned?

For RQ 2, a random forest classifier was built to predict whether a student learned, from mid-to post-test using COPES variables. For that classifier, we split students into a group that learned (NLG>0) and a group that did not learn (NLG<=0). Data were partitioned into a 75:25 split for training and testing. To strike a balance between the best AUC and the lowest...
error rate, grid search was used for the 1) number of trees grown and 2) COPES features used (Table 1) to grow each tree. Ten repetitions of 10-fold cross-validation (in the training) were used. The random forest classifier achieved an AUC ROC of .63 (95% CI [.54-.72] and Cohen’s kappa of .2553, with 2 features per decision tree. The important features for classifying NLG were, first, conditions, general programming prior knowledge, and second, perceived programming ability. The third most important feature was the evaluations students used during programming, followed by the operators used to debug non-compiler errors. The operators used to debug compiler errors had a mean importance of 0.

4. Discussion

Novice programmers do not utilize SRL effectively with automated feedback tools in CS education (Ko et al., 2019; Loksa et al., 2022; Marwan et al., 2022). In this paper, we explored the role of prior knowledge on SRL and its relation to learning outcomes. The first RQ found that novices used operators more to debug compiler errors on failed programs. In contrast, more experienced programmers used more operators to debug non-compiler errors. By contrast, the groups did not differ in the evaluations of failed programs.

The second RQ found that the most important COPES feature for predicting students’ learning gain was their level of general programming knowledge prior to the course. This was followed by their perceived programming ability, how often they evaluated failed programs with the automated feedback tool, and use of operators to debug non-compiler errors. The number of operators used to debug compiler errors was not part of the predictors of students’ learning. Although syntax plays important role at the initial stage, its importance decreases over time. This work advances our understanding on the role of prior programming knowledge on SRL with an automated feedback tool and its relation on learning outcomes in CS education. This research has two limitations. First, the use of log data to study COPES, possibly missing other COPES behaviors that occur beyond system interactions, and 2) products were not collected, possibly missing additional data on SRL during programming.

References


Construction of a Japanese Language Learning Support System for Learning Semantic Negotiation

Satoru Kogure\textsuperscript{a*}, Akira YOSHIDA\textsuperscript{b}, Yashihiro NOGUCHI\textsuperscript{a}, Koichi YAMASHITA\textsuperscript{c}, Tatsuhiro KONISHI\textsuperscript{a} & Makoto KONDO\textsuperscript{a}

\textsuperscript{a}Faculty of Informatics, Shizuoka University, Japan
\textsuperscript{b}Graduate School of Integrated Science and Technology, Shizuoka University, Japan
\textsuperscript{c}Faculty of Business Administration, Tokoha University, Japan
* kogure@inf.shizuoka.ac.jp

Abstract: Semantic negotiation techniques are important for beginners of Japanese learner. In this study, we investigated how semantic negotiation is handled in a learning environment and then reclassified it. We developed a Japanese language learning support system that can handle semantic negotiation using text input. We report the results of a questionnaire survey administered to three Japanese language teachers and an experimental evaluation of a foreign student.

Keywords: Semantic negotiation, Japanese language learning, role-play, cooperative learner agent

1. Introduction

The process of interaction between participants in a conversation, such as clarification requests and confirmation checks, to understand the intentions of the other party, is called semantic negotiation (SN) (Varonis, & Gass, 1985; Long, 1996). SN is important for second language (L2) learners, who are likely to experience frequent communication problems in the language they are learning. However, Miyazaki pointed out that in Japan, "research results on SN are not reflected in Japanese language education" (Miyazaki, 2020). In fact, our survey of instructional materials used in Japanese language education (Three A Network, 2016) revealed that most students did not use SN. We also administered a questionnaire survey to Japanese language teachers. We found that three out of the four teachers did not teach SN in their classes because it was not a part of the curriculum. Thus, SN is not considered important in the field of Japanese language education. To solve these problems, we constructed an environment that enables self-taught learners to study SN in Japanese.

The objectives of this research are (1) to propose a learning scenario that enables beginner learners of Japanese to learn SN as a means of adjusting to problems that occur in dialogue, and (2) to construct a text-based Japanese dialogue system that helps students to learn dialogue and SN in Japanese through conversation-based experiences with a cooperative learner agent (CLA).

2. Semantic Negotiation

The classification of SN in Japan has not been well established (Yokobayashi, 1991; Fujinaka, 1996; Igari, 1998). In this study, we newly categorized and defined SN with reference to previous studies that deal with the classification of SN. The categories of SN defined in this study can be divided into those conducted by the learner and those conducted by partner. SNs for learner include (A) reduction strategy and (B) achievement strategy. (B) strategy can be divided into (B1) speech strategy and (B2) listening strategy. In this study,
we focused on (B1) speech strategies and classified them in detail. (B1) speech strategies can be divided into (a) direct appeal, (b) indirect appeal, (c) non-linguistic, (d) implicit signal of uncertainly, (e) confirmation of partner's understanding, (f) code switching, (g) literal translation, (h) paraphrase, (i) substitution, (j) generalization, (k) neologism and (l) restructuring. SNs for partner include (i) error corrections, (ii) contributors/completions, (iii) elaborations, (iv) confirmation checks and (v) clarification requests. Among the negotiation of meaning conducted by the learner, (A) the "reduction strategy" is not dealt with because it is difficult for the system to recognize it. In this study, we primarily addressed (B) "achievement strategies". Among (B) "accomplishment strategies," (B2) the "listening strategy" is limited to cases in which the system is not sure of the meaning of the other party's utterance and the user wants to confirm whether his/her understanding is appropriate, according to the specification of the system for text-based dialogues. Therefore, (B2) "listening-back strategy" is a SN, in which the learner responds to the other party's utterance with a question, and the other party responds to the SN by answering the question. For (B1) "speech strategies," this study covers SN (a), (e), (f), (g), (h), (i), (j), (k), and (l). We did not use SN (b), (c), and (d) because they cannot be realized in a text-based learning support system.

To investigate how much SN is actually taught in the field of education, we administered a questionnaire survey to four teachers who teach Japanese at Shizuoka University. In terms of learning materials, one respondent used "dialogue including SN without explicitly stating it" while another respondent used "dialogue including SN with explicitly stating it" as a learning material, and two respondents did not use "dialogue including SN." This indicates that SN is not always taught in Japanese language study. In the classes that used "dialogue including SN without explicitly stating it," only (i) error corrections and (iv) confirmation checks were used, and the learners' (B1) "speech strategies" were not covered. On the other hand, the lessons that involved "dialogue that includes SN" as the learning material dealt with (A) reduction strategies, (B1) speech strategies, and (B2) strategies for listening back (i) "corrections," (ii) "contributions and completions," (iv) confirmation checks, and (v) clarification requests. When focusing on the "speech strategies," only one of the four respondents used the material including these strategies in their classes. Two respondents who do not employ SN stated that they do not teach it because it is not part of the curriculum. This indicates that SN, or speech strategies, are not taught in Japanese language learning.


In this study, we constructed a text-based dialogue system that supports SN and allows for free input. The system is implemented in three parts: (1) scenario selection, in which the learner chooses a scenario to be learned with reference to learning goals, learning content, and dialogue situations; (2) a dialogue portion, in which the learner interacts with the agent according to the situation provided; and (3) feedback, in which the system evaluates the learner's dialogue from the teacher's perspective. In this system, in order to "set up a lesson specialized for the learning of a certain SN," expressions that trigger the detection of the use of that negotiation are registered in advance in the contents prepared by the teacher. The number of SN that a learner can learn acquire simultaneously is limited to one; hence, there is no error in the detection of SN. The teacher needs to prepare two pieces of content in advance for each lesson: "situation contents" and "SN contents". "Situation content" includes the image file name of the situation, the caption, the learner's assumed utterances and the CLA's utterances, as well as the results of the dependency parsing of the learner's utterances. SN content includes the type of SN to be studied, the words and phrases that trigger the negotiation, the learner's sentences assumed used in the negotiation, and the results of the analysis of the engagement of these sentences.
4. Experimental Evaluation

To evaluate the usability of the system, we asked three Japanese language teachers who teach Japanese to foreign students and one foreign student who is studying Japanese to use the system. In an experiment with three teachers, we first (1) explained SN to the teachers, (2) asked them to view slides and videos on how to use the system, and (3) asked them to view slides and videos on how to create content for this system. From the results of the questionnaire, we obtained a rating (3.00 out of 5) that the system is somewhat effective in helping students learn SN. On the other hand, we received 2.67 and 2.33 in the “would you want your students to use it” and “would you use it in class” categories, respectively. These results indicate that there is room for improvement of the system. In the experiment with one international student, we administered (1) a pre-questionnaire, (2) an explanation of SN and system usage, (3) a tutorial on system usage, (4) the first learning session of SN, (5) the second learning session of SN, and (6) a post-questionnaire. This foreign student had not learned SN explicitly in class, but was already familiar with SN itself. We received a score of 4 and a score of 5 (on a 5-point scale) in the evaluation of whether the system was useful for learning SN and whether we would like to use the system in the future, respectively.

5. Conclusion

We developed a Japanese language learning support system that enables self-study of SN, based on the fact that SN techniques are very important for beginners in Japanese, but have not been adopted in the broader field of Japanese language education. The results of the questionnaire survey of two Japanese language teachers suggested that the system itself is effective, but when asked if they would want to use it in their classes, they gave it a score of three or less out of five. In the future, it will be necessary to conduct experimental evaluations with a larger number of Japanese language learners and to develop an authoring tool that enables teachers who are not familiar with IT systems to create content.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K12172.

References


Creating Meaningful Connections: The Role of Simultaneous Multi-Situational Learning in Knowledge Contextualization and Application

Chih-Yen CHEN\textsuperscript{a}, Su-Hang YANG\textsuperscript{b}, Meng-Xuan XIE\textsuperscript{a}, Yi-Chuan FAN\textsuperscript{c}, Jen-Hang WANG\textsuperscript{c}\textsuperscript{*} & Gwo-Dong CHEN\textsuperscript{a}

\textsuperscript{a} Department of Computer Science and Information Engineering, National Central University, Taiwan
\textsuperscript{b} Department of Hospitality Management, Chien Hsin University of Science and Technology, Taiwan
\textsuperscript{c} Research Center for Science and Technology for Learning, National Central University, Taiwan
\textsuperscript{*}harry@cl.ncu.edu.tw

Abstract: This study investigates the effectiveness of a simultaneous multi-situational learning system with embodied recognition to enhance learners’ outcomes. Traditional situational learning systems provide a single situation for knowledge explanation or application, which limits learners’ ability to fully immerse themselves in the learning environment and receive immediate feedback. To address this, a simultaneous multi-situational learning system was proposed to create a digital space with multiple situations, allowing learners to immerse themselves in different scenarios simultaneously and facilitating reflective learning. An experimental study with 60 undergraduate students in a Hospitality Japanese course compared a control group using the traditional situational learning system with an experimental group using the proposed simultaneous multi-situational learning system. Results showed the proposed system significantly improved learning outcomes, understanding, and interest in learning. It enables the prompt creation of situational learning environments, contextualizes knowledge, and fosters reflective learning with real-time feedback. Overall, this study contributes to the field of education by providing a more effective approach to situational learning that can be easily implemented in classroom settings.

Keywords: Situational learning, embodied cognition, Simultaneous multi-situation

1. Introduction

1.1 Background

Situational learning is an approach that allows learners to immerse themselves in situations where knowledge may be applied, thereby strengthening the connection between knowledge and its practical applications in real-world situations. Brown, Collins, and Duguid (1989) emphasized the importance of learning in an environment that corresponds to its situations and showed that knowledge can be contextualized. Consequently, researchers explored how to design a situational learning environment to enhance learners' learning outcomes, and they showed that situational learning is an effective method helping learners to acquire knowledge (Herrington & Oliver, 2000). On the other hand, creating an appropriate situational environment in the real world can be time-consuming and complicated, making situational learning rarely be applied in traditional classrooms (Dede, 2009).

However, with the development of technologies, virtual environments can be constructed more easily, leading the provision of situational learning methods in virtual environments. Wu et al. (2015) have proposed an approach in which real human images are combined with virtual scenes to enable learners to not only immerse themselves in the situation but also see their own appearance and inspect performance in the virtual
environment. This approach can be particularly useful in establishing a situational learning environment promptly in classrooms and facilitating reflective learning.

In addition, embodied cognition theory shows that learners can acquire knowledge through the interaction between their bodies and the environment (Wilson, 2002). Furthermore, some research suggested that embodied interaction in virtual space can improve user engagement (Lindgren et al., 2016). Recently, Wang et al. (2020) demonstrated that the integration of body recognition services in digital learning environments allows for learners to interact using embodied behaviors. Furthermore, computers can function as personal instructors, identifying inadequate aspects of posture and emotional expression during the learning process, and guiding learners to adopt correct behaviors (Xie et al., 2022).

However, during a learning process, the context is not limited to only one situation. The Mantle of the Expert (MoE) theory, proposed by Heathcote and Bolton (1994), suggests that when learners are asked to present a professional role by the instructor, they must learn relevant knowledge and imagine a situation to achieve their goal. As a result, they will need to play a variety of different roles when they face different events that change at the same time. Additionally, the Legitimate Peripheral Participation (LPP) theory (Lave & Wenger, 1991) states that the best way to learn complex skills is to become a member of a problem-solving team. Therefore, it is necessary to understand the status of different roles in different situations. As demonstrated, multiple scenarios are essential to explain various situations during situational learning.

1.2 Motivation

As aforementioned, it is evident that current studies typically restrict learners' learning to a single situation in which they only witness the application of knowledge but do not fully grasp how it is contextualized. However, with the advancements in technology, particularly in the fields of AI, GPU, memory, and related technologies, situational learning systems can now be combined with recognition services. This integration facilitates the provision of situational feedback by detecting learners' embodied behavior while simultaneously presenting multiple situations. Consequently, learners can immerse themselves in multiple scenarios simultaneously, enabling them to contextualize and apply their knowledge in various contexts.

1.3 Research Objectives

This study aims to build a learning system that can identify users' embodied behavior and provide them with the situation of knowledge explanation, the situation of knowledge application, and the situation of post-problem-solving. The main goal is to verify whether this system can improve learners learning performance by allowing them to immerse in simultaneous multi-situational scenarios.

1.4 Research Question

This study investigates whether the proposed simultaneous multi-situational learning system, which generates different situations for the needs of knowledge learning and service and enables the coexistence of various situations in a digital space, can lead to better learning achievement compared to traditional single-situation learning systems. Therefore, the research question is: "Could the simultaneous multi-situational learning system outperform traditional single-situation learning systems in terms of learning achievement?"

2. Related Work

2.1 Situational Learning
Situational learning is a valuable approach that emphasizes the interaction between learners and the environment to acquire knowledge. Brown et al. (1989) emphasized that knowledge is acquired through interaction with contexts of the environment, making situational learning a valuable approach to strengthen the connection between knowledge and practical application. Gagne et al. (2005) also argued that learning is most effective when it is applied to real-life situations. To improve learners' thinking, exploration, and problem-solving abilities, situational learning aims to immerse learners in various situations for exploration and reflection. Scholars have proposed various methods of situational learning, including authentic learning and drama-based learning. Authentic learning combines learning content with real-world issues, allowing learners to learn and practice at the same time, which has shown to improve learning outcomes (Ornellas et al., 2018). Drama-based learning involves learners collaborating to perform a drama, which can increase learner engagement and promote active learning (Dawson & Lee, 2018).

Additionally, previous research on situational learning has used different approaches, such as situational classrooms and simulated environments. The English Village is an example of a situational classroom that transforms a traditional classroom into a foreign language learning environment (Kelch, 2011). Similarly, the Clinical Skills Center of the National Taiwan University Hospital has created simulated spaces for performing surgeries and medical examinations to allow learners to become familiar with the process of seeing a doctor. With advances in technology, situational learning has been combined with multimedia technology to simulate learning situations in real environments. For instance, Chou et al. (2012) developed a computer game that combines game-based learning with situational learning, enabling elementary school students to learn how to react during earthquakes. Moreover, digital spaces built in the classroom based on drama-based learning enable learners to perform dramas and learn in the situation (Wu, 2015). All of the above confirm the feasibility of digital learning platforms for situational learning.

2.2 Embodied cognition

Recent technological advancements have made it possible to create realistic environments in a digital way. Studies have indicated that situational learning in digital reality can improve learning outcomes (Giasiranis & Sofos, 2016). In recent years, human-machine interaction has become more diverse, with additional sensors or recognition systems detecting users' embodied behaviors. Linden Lab's Second Life, for instance, uses Windows, Icon, Menu, and Pointer (WIMP) for interaction, allowing learners to control virtual avatars through devices and interact with the virtual world through WIMP. With the development of VR and AR, users can experience more immersive interactive experiences (Buttussi & Chittaro, 2017). Research suggests that using VR in education can enhance learners' learning motivation and improve their performance, outperforming observation and exploration (Kavanagh et al., 2017; Kwon, 2019).

Situational learning combined with embodied interaction can be advantageous for educational applications. However, in most embodied interaction situations, users enter the digital world as a virtual avatar, which makes it difficult for them to observe their own performance in the digital world and engage in reflective learning. To address this limitation, Wu et al. (2015) developed the Digital Learning Theater (DLT), a drama-based digital reality system that allows learners to immerse themselves in the digital environment through a mirror image, enabling them to see their own performance during the drama and engage in reflective learning.

3. Implementation

3.1 Design Concept
In this study, we designed a simultaneous multi-situational learning system that aims to enhance learners' connection between situations and knowledge. The system consists of four situations: the "situation of knowledge presentation," "situation of knowledge explanation," "situation of knowledge application," and "situation of post-problem-solving." The "situation of knowledge presentation" provides learners with the necessary concepts and knowledge through textbooks and lectures. The "situation of knowledge explanation" is an interactive mechanism that allows learners to select specific contents for more in-depth learning. The "situation of knowledge application" provides learners with scenarios that allow them to apply what they have learned to real-world situations. Finally, the "situation of post-problem-solving" provides a feedback mechanism that allows learners to reflect on and incorporate the knowledge and experience gained from the previous situations into their own learning process. To implement the system, learners will access the different situations through a digital platform, and each situation will be designed to provide a unique learning experience. The system will evaluate learners' performance by monitoring their embodied behavior during each situation and providing feedback in the "situation of post-problem-solving." The feedback will be designed to help learners reflect on their performance and incorporate their learning into future situations, thereby improving their learning outcomes.

3.2 System structure

The system design of this study consists of three modules: the drama-based simultaneous multi-situational display system, the simultaneous multi-situational editor, and the database, as shown in Figure 1. Instructors and learners can use the simultaneous multi-situational editor to manage materials and edit the script stored in the database. When learners perform in a drama, the simultaneous multi-situational display system retrieves the relevant data from the database and overlays multiple scenes, changes scenes, and dresses focus situations. Finally, the learners can observe their appearance and performance in the digital space via the displayed video.

![System structure](image)

The simultaneous multi-situational learning system comprises three parts. The first part, shown in Figure 2, is the data editing system, which allows editors to add events, materials, scenes, props, and costumes that occur when the event is triggered. The second part is the embodied situational learning system, which processes the events set in the
database according to the previous editing system. This system mainly involves overlapping multiple scenes, scene changes, and focus situation dressing. Additionally, the system recognizes and processes the generated scene and the original situational scene according to the learner's action response. The third part is the database, which stores the learning materials and scripts based on the Hospitality Japanese textbooks. The teaching content includes the complete reception process of the waiters, starting from the reception, menu introduction to the checkout. These materials are discussed with the instructors of the Hospitality Japanese course to help learners learn Hospitality Japanese and the response actions during the reception process.

3.3 Real-time embodied cognitive system

The real-time recognition cognitive system consists of two parts, including speech recognition and body recognition. Figures 3 and 4 illustrate the flowchart of the system.

Speech recognition is the first part of the system, which is processed through the speech recognition service in Microsoft Azure. The Speech-to-Text API converts audio into Japanese text, and punctuation processing is performed to minimize errors during content comparison. The processed results are then compared with the lines, enabling the system to make a corresponding situational change to the background and props based on the recognized results.

The second part of the system is body recognition, which involves training the recognition model using TensorFlow and the Long-term Recurrent Convolutional Networks (LRCN) method. This model combines CNN and LSTM, where Convolutional layers extract features from each video frame, and these features are sent to the LSTM layers to process the time relationship between each frame and improve recognition of continuous actions. This model solves the issue that traditional CNN models can only recognize a single frame and simulates the impact of the user's performance on the situation. After identifying the body
recognition results, the system makes a corresponding situational change to the background and props based on the recognition results.

*Figure 4. Body recognition flowchart*

4. Method

4.1 Experimental Hypothesis

To evaluate the effectiveness of the proposed system, the present study tests the following hypothesis: The use of a simultaneous multi-situational learning system that provides immediate feedback based on learners' embodied interactions and simultaneously presents multiple situations is more effective in terms of learning achievement than a traditional situational learning system that only provides a single situation. The study aims to compare the learning outcomes of the two systems to determine whether the proposed simultaneous multi-situational learning system outperforms traditional situational learning systems in terms of learning achievement.

4.2 Experimental Subject

This study was conducted in collaboration with a local university of science and technology located. The sample consisted of 60 participants from two classes in the Department of Hospitality Management. The two classes were randomly assigned to either the experimental group (30 participants) or the control group (30 participants). To ensure comparability between groups, participants within each class were divided into small groups of six based on their pre-test scores. The experimental group was exposed to a simultaneous multi-situational learning system that provided additional situational changes based on learners' embodied behaviors during learning activities, while the control group received a traditional single-situation learning system where the instructor changed the situations on the screen according to learners' embodied behaviors during the rehearsal of the drama.

4.3 Experimental Process

This study employed a six-week experimental process, with one class session per week and two hours per class session, which is depicted in Figure 5. The instructor for the two groups was the same.
During the first week, the learners took a pre-test and received an introduction to the system. Over the next two to four weeks, the course content was taught, and group practice sessions were conducted. In the fifth week, the learners performed in a formal setting, and a post-test and questionnaire were administered during the final week. To enhance the learning experience, the system was also accessible online, providing learners with the opportunity to practice on their own time and increase their familiarity with the system.

The environmental design of curriculum is shown in Figure 6, where the classroom is divided into an actor area and an audience area. The screenplay in the casting area differs between control group and experimental group, where the experimental group used the simultaneous multi-situational learning system, while the control group used a traditional single-situation learning system. The control group only depicts the knowledge application situation in the restaurant scene, while the experimental group presents a multi-situational screenplay combining two situations: the situation of knowledge explanation for the description of ingredients and the situation of knowledge application for the description of the restaurant scene.

4.4 Experimental Measurement Tools

To evaluate the effectiveness of the proposed simultaneous multi-situational learning system, pre-tests and post-tests were conducted to assess the learners' academic knowledge in hospitality situations, their proficiency in Japanese oral communication, and their familiarity...
with Japanese culture. These tests were based on the Hospitality Japanese textbook used in
the course and were designed in consultation with the course instructor. The tests included
true/false, multiple-choice, and fill-in-the-blank questions, with a maximum score of 100 for
each test. Analysis of Covariance (ANCOVA) was employed on the test scores to analyze the
impact of the learning system on the learners' academic performance. Additionally, a
questionnaire was administered using a Likert five-point scale to understand the learners' perceptions on the system's impact on their motivation and confidence. The questionnaire was designed based on the ARCS model (Keller, 1987) and was validated by expert professors, with a Cronbach's alpha of 0.972, indicating high reliability.

5. Result

5.1 Pre-Test and Post-Test Result

To evaluate the normality of the data, Shapiro-Wilk test was employed for both the experimental and control groups as the sample size for each group was less than fifty. The results showed that the experimental group pre-test ($p = .120 > .05$), control group pre-test ($p = .145 > .05$), experimental group post-test ($p = .111 > .05$), and control group post-test ($p = .432 > .05$) all fulfilled the assumption of normal distribution. To further test the assumption of homogeneity of regression coefficients, the results showed that the assumption was met. ($F = .015, p = .902 > .05$). Moreover, the assumption of the equality of variances was also supported by the result of Levene's test ($p = .470 > .05$).

Based on the above results, ANCOVA was performed. The result shows that the average and the adjusted mean of the experimental group post-test were 73.673 and 72.293, respectively; while those of the control group were 65.507 and 66.257, respectively. Furthermore, there is a significant difference between the two groups ($F = 4.303, p = .043 < .05$). The partial eta square value was .07, with a range from .06 to .14, which indicates the effect size is at a medium level (Cohen, 1992). The data indicate that the simultaneous multi-situational learning system was significantly more effective than the traditional situational learning system with a single situation in terms of learning performance.

Therefore, the results of this study support the hypothesis that the use of a simultaneous multi-situational learning system is more effective than using a traditional situational learning system with a single situation.

5.2 Questionnaire Result

The results of the questionnaire indicated that there were no significant differences in the mean scores of the user's system experience, learning motivation, and building confidence between the experimental and control groups. Nonetheless, both groups provided positive feedback on the experimental content, demonstrating that multimedia tools can effectively engage and motivate learners' interest, as well as align with the principles of situational learning and enhance their motivation to learn.

5.3 Interviews

In order to gain a deeper understanding of the experimental group's experience, interviews were conducted with four learners and the instructor who participated in the experiment. In terms of system control, the participants expressed that the simultaneous multi-situational learning system provided a novel control method. Compared to using mobile phones or computers, this method provided a more immersive experience. Some learners also expressed a desire for the system to point out their errors during learning, which could improve their long-term learning performance.

Regarding situational changes, learners indicated that the application of the situation, explanation of the situation, and the resulting situation after the change can effectively link
them to the knowledge they want to express. The learning model allowed them to have a deeper understanding of what they were learning and a better understanding of the textbook knowledge. Furthermore, the learning model increased their motivation to learn. Some participants stated that the system had made them eager to express more diverse knowledge while others hoped for stronger audio or visual effects.

6. Conclusion

The aim of this study was to investigate the effectiveness of a simultaneous multi-situational learning system on learners' learning achievement. The proposed system created a digital space with multiple situations, allowing learners to perform and reflect on their behavior in real-time. This approach addressed the limitations of traditional single-situation learning, such as the difficulty of immersing learners in a single situation and the lack of real-time assessment. The experimental results demonstrated that the performance of the experimental group was significantly better than that of the control group in the post-test.

Furthermore, the questionnaire and interview results indicated that the digital space with multiple situations enhanced learners' understanding and connection to the knowledge and increased their interest in learning. Therefore, the contribution of this study is the proposal of a new type of simultaneous multi-situational learning system and the validation of the hypothesis that a simultaneous multi-situational learning system is more effective than a traditional situational learning system.

Based on the feedback from the experimental group's interviews, students expressed that some variations in situations were not engaging enough, future research could focus on exploring the incorporation of sound effects to enhance the overall attractiveness and effectiveness of the learning experience. Furthermore, with advancements in Augmented Reality (AR) and Mixed Reality (MR) technologies, such as the Mixed Reality Toolkit (MRTK) offered by Microsoft, visual effects could be enhanced to create more realistic multi-situational environments. Moreover, adding error guidance could further enhance the learning process, allowing learners to instantly correct their posture, pronunciation, facial expressions, and other aspects of knowledge that are not typically expressed through traditional paper-based learning.

Acknowledgements

This study is supported by the National Science and Technology Council, Taiwan. Grand Number: MOST 109-2511-H-008-004-MY3 and NSTC 112-2811-H-008-006.

Reference


ECLAIR: A Centralized AI-Powered Recommendations System in a Multi-Node EXAIT System

Isanka WIJERATHNEa, Brendan FLANAGANb, Yiling DAIa & Hiroaki OGATAa

aAcademic Center for Computing and Media Studies, Kyoto University, Japan
bCenter for Innovative Research and Education in Data Science, Kyoto University, Japan

*Wijerathne.isanka.6z@kyoto-u.ac.jp

Abstract: Educational recommender systems are increasingly becoming a core feature of modern educational systems. Often the recommender component of a system is tightly integrated, or might be remotely located without accessing data from other local systems. This paper proposes a framework called ECLAIR in which local educational systems can work and share data with a global recommender system that spans multiple educational institutions. In particular, an AI-driven recommendation system is intricately integrated within a multi-node learning management system. Situated at the intersection of large-scale data analysis and personalized education, ECLAIR efficiently processes heterogeneous data from diverse LMS, while ensuring data security and privacy. The proposed ECLAIR's architecture, data pipeline, and processing mechanisms are explored in detail, focusing on its seamless integration with the existing infrastructure. By leveraging MongoDB change streams and relational databases, ECLAIR guarantees real-time data synchronization, secure storage, and efficient processing. Its unique Ingestor tool transforms selected xAPI data points into a relational table format, bolstering system functionality. The successful launch of ECLAIR serves as a testament to the potential of AI in enhancing personalized learning experiences, improving data security, and bolstering system efficiency. Nevertheless, the paper emphasizes the need for ongoing research, specifically concerning privacy-preserving mechanisms and efficient management of data heterogeneity. It demonstrates ECLAIR's pivotal role in the rapidly evolving landscape of eLearning, and its potential for future advancements, scalability, and adaptability, setting a new precedent for AI-powered eLearning recommendation systems.

Keywords: Artificial intelligence (AI), AI-Powered recommender system, Explainable AI, e-learning, ECLAIR, EXAIT

1. Introduction

In an era marked by technological integration, education has witnessed considerable transformations. The emergent realms of Learning Management Systems (LMS), eLearning platforms, and Artificial Intelligence (AI)-enhanced recommendation systems have significantly augmented educational experiences (Aleven et al., 2016). These advancements demonstrate potential in personalizing learning experiences, optimizing data security, and augmenting system efficiency. However, despite these gains, obstacles persist, such as managing data heterogeneity, preserving data privacy, and maintaining system scalability (Boitshwarelo, 2011). This presents a continued need for innovative approaches.

In this paper we propose the EXAIT CentraL Artificial Intelligence Recommender (ECLAIR), a centralized AI-powered recommendation system that functions within a multi-node EXAIT (Educational eXplainable AI Tools) system (Flanagan et al., 2021). ECLAIR is designed to tackle the key challenges associated with AI recommender systems in eLearning, offering a novel pathway towards personalized learning.

The ECLAIR system stands at the intersection of large-scale data analysis and
personalized education. It can manage and analyze a vast array of data across various LMS platforms, furnishing personalized, efficient, and precise recommendations to learners. ECLAIR prioritizes data security and privacy, crucial considerations in the global educational sector (Polonetsky & Tene, 2013; Wijerathne et al., 2022).

The paper is methodically structured to provide an extensive understanding of the ECLAIR system, commencing with a review of related works in the field of AI-enhanced recommendation systems. This is succeeded by a comprehensive system overview, an explanation of the data pipeline and processing mechanisms, and a detailed account of the implementation and integration of ECLAIR. Throughout the paper, we underscore how ECLAIR surpasses traditional system challenges and anticipates future advancements in the rapidly evolving domain of eLearning.

2. Related Works

The intersection of Learning Management Systems (LMS), eLearning, and Artificial Intelligence (AI) has been a focal point of research in recent years (Ouyang et al., 2022), particularly concerning the development of recommendation systems. One of the earliest strides in this space was the adoption of collaborative filtering and content-based recommendation algorithms (Koren, 2008; Lu et al., 2015; Marlin, 2003). For instance, implemented a personalized recommendation system in eLearning platforms using a content-Lalitha et al. (Lalitha & Sreeja, 2020) based approach, which considered the learners’ profiles and their learning styles. Despite the initial success, traditional methods like these often-encountered challenges such as the cold start problem and scalability issues, particularly when dealing with a vast number of users and contents (Bell & Koren, 2007; Lam et al., 2008; Lika et al., 2014).

To overcome these hurdles, subsequent research began to incorporate machine learning techniques into the design of recommendation systems. A study conducted by Huang et al. (Huang et al., 2002) exemplifies this transition. They proposed a hybrid recommendation system that combined the strengths of collaborative filtering and deep learning techniques. The system was built on an Autoencoder, a neural network model that improved the accuracy of recommendations by learning complex, non-linear user-item interactions. This study represented a substantial step forward from traditional methods, demonstrating the potential of machine learning in enhancing recommendation systems (Bai et al., 2019; Beel et al., 2013).

Numerous educational institutions traditionally exhibit a reluctance to disclose student learning data to external entities, a tendency grounded in concerns regarding privacy and data security. However, the functionality and effectiveness of recommender systems are inextricably linked to their access to extensive datasets, thereby creating a conundrum. As such, there is a pressing need to strike a delicate equilibrium between data utility, that is, the optimization of recommender systems through comprehensive data access, and the preservation of data privacy, ensuring the protection of sensitive information. This balance is paramount to harness the full potential of such systems while still adhering to ethical guidelines and legal regulations (Kobsa, 2007; Polonetsky & Tene, 2013). Ocheja et al's work introduces the Blockchain of Learning Logs (BOLL), a platform that comprehensively records learning pathways and achievements beyond the scope of traditional transcripts. This system collects data from distributed learning systems and addresses the “cold-start” issue in learning data analytic platforms by enabling the secure and verifiable transfer of learning records across institutions. Crucially, BOLL maintains a delicate balance between data privacy, sensitivity, and technology, underscoring its commitment to respecting individual privacy (Ocheja et al., 2019).

In the preliminary phase, the proposed artificial intelligence (AI)-driven recommendation system has been utilized for teaching English as a Foreign Language (EFL) to high school students (Taki et al., 2022). Furthermore, an AI-enhanced recommendation system with self-explanatory capabilities has been used for mathematics instruction (Dai et al., 2022; Flanagan et al., 2021).
These advancements underscore the potential of centralized AI-powered recommendation systems in enhancing personalized learning, improving data security, and increasing system efficiency. However, they also highlight the need for additional research in areas like privacy-preserving mechanisms and managing data heterogeneity.

![Figure 1: Overview of the proposed system](image)

3. System Overview

The implementation of AI recommender systems, which inherently require substantial amounts of data for efficient operation, is thereby affected. It is pertinent to understand that the quality of these systems is directly proportional to the quantity of data available, hence a larger dataset is always advantageous.

In response to this challenge, we propose a system that amalgamates the requisite data from these distributed systems into a singular, centralized system as shown in Figure 1. The proposed system contains two main parts: Sets of multiple LEAF systems (Left side) and the central AI recommender (right side), with API connections facilitating communication between the two systems. This system then processes the accumulated data through the AI recommender system. The processed information, in turn, is channeled back to the respective LEAF system, thereby ensuring a feedback mechanism that enhances the overall learning experience.

3.1 LEAF System

![Figure 2: Overview of the LEAF system](image)

The Learning Evidence Analytics Framework (LEAF) infrastructure shown in Figure 2 (Flanagan & Ogata, 2018) is an intricate composite of several unique components, prominently including an LMS, a digital learning material reading systems Bookroll, a learning analysis tool Log palette, and a Learning Record Store (LRS). LRS, is an open-source tool for managing learning data, leveraging the scalable and high-performance capabilities of MongoDB for data storage and retrieval, and supporting data analysis and insights generation. The LMS, as implied by its nomenclature, orchestrates the planning, execution, and evaluation
of the comprehensive learning process. Meanwhile, Bookroll functions as a versatile e-book reader. Intriguingly, all these systems emit Experience API (xAPI) data, which are subsequently stored within the LRS. The proposed system has multiple LEAF nodes connected to the Central AI Recommender.

3.2 Central AI Recommender

The Central AI Recommender as shown in Figure 1 is a complex construct consisting of numerous key components such as the Recommender Model Trainer, the LRS Database, the LP Database, the BR Database, the AI Recommender Database, the Ingestor, and the AI Recommender application. Data harvested from the distributed LEAF systems through the Egress API is consolidated within the LRS, LP, and BR databases. This stored data is subsequently leveraged by the Recommender Model Trainer to train the model, which is then deposited in the AI Recommender Database. The Ingestor serves as a data transformer, converting LRS MongoDB data into relational data, which is then stored in the AI Recommender Database. Finally, the AI Recommender application utilizes this data to provide recommendations through the Kodama API.

4. Data Pipeline and Processing

The LEAF system encompasses two types of data: relational and non-relational. Relational data are managed through the MySQL database and are primarily utilized within the LMS, Bookroll, and Log Palette. In contrast, non-relational data, mainly characterized by xAPI statements, are generated from every interaction within these systems and are subsequently stored in MongoDB. The proposed system, moreover, features multiple LEAF nodes distributed across various educational institutions.

The Experience API (Tin Can API1), initially introduced in 2013 (Gonzalez & Churchill, 2019), serves as a crucial component of learning analytics. It offers a platform-agnostic formalism for capturing events that transpire throughout any learning experience. The proposed system capitalizes on the capabilities of xAPI to record all learner-system interactions. The generated xAPI statements are committed to a Learning Record Store (LRS), which is a key element in the infrastructure for storing and retrieving learning records. This combination of relational and non-relational data storage mechanisms, coupled with the adoption of xAPI for event tracking, ensures comprehensive data coverage and fosters the potential for rich analytical insights.

For the purpose of enhancing the ECLAIR system, it is imperative that pertinent data is obtained from various LEAF data stores, which will subsequently be utilized in training its models. This process is essential to avoid congestion and to maintain a keen focus on the microservices architecture. In light of this, a system has been devised that incorporates the LEAF Data Layer (LDL). This system comprises two Application Programming Interfaces (API) endpoints, named "Egress" and "Kodama", with the later meaning echo in Japanese. These designated endpoints are instrumental in facilitating seamless data exchange between the LEAF systems and the ECLAIR.

Consequently, the ECLAIR system has been provisioned with the ability to subscribe to the LRS MongoDB databases of the LEAF systems. This subscription operation is actualized by leveraging MongoDB change streams, a feature that facilitates real-time data processing. This mechanism empowers the ECLAIR system to garner real-time data from the LRS MongoDB, and in turn, replicate this data into its own LRS MongoDB, thereby ensuring a continuous flow and synchronization of data between systems.

Moreover, the ECLAIR system is designed to extract relational data from the LEAF systems through the LDL EGRESS API. The extracted data is then transferred and securely stored within the ECLAIR's own data repositories. This methodical process ensures that the ECLAIR system continuously has access to updated and relevant data necessary for its

1 https://xapi.com/
optimal operation and the achievement of its intended objectives.

The ECLAIR necessitates relational data for the comprehensive functionality and features it offers. To achieve this, selected data points from the replicated LRS MongoDB data are transformed into a relational table format using the Ingestor tool. The Ingestor system is comprised of two main components: real-time data conversion and backfilling data. In scenarios involving system failure or the incorporation of novel xAPI data points, the backfilling application can be manually executed to facilitate the propagation of relational data from the LRS MongoDB.

5. Implementation and Integration

During the implementation phase, our focus was on integrating the ECLAIR system with the existing LEAF infrastructure. This integration, which was aimed at facilitating seamless data transmission, synchronization, and processing while also guaranteeing data security and privacy, necessitated meticulous planning and stringent execution.

The ECLAIR system is based on a microservices design, which allows for versatility, scalability, and simplicity of updating. It was created and built mostly with Python, a versatile and widely used programming language for AI applications, and the Scikit-Learn\(^2\) module, which provides a wide range of machine learning techniques for the recommendation.

Setting up the Egress and Kodama endpoints for data interchange between the LEAF system and ECLAIR was the first step in its implementation. These endpoints were built with the Flask framework, which allows for the creation of RESTful APIs. Following that, the ECLAIR system was set up to use MongoDB change streams to subscribe to the LEAF system's LRS MongoDB databases. This subscription facilitates real-time data processing and synchronization between systems. The relational data obtained from the LEAF systems through the LDL Egress API was incorporated into ECLAIR's own data repositories. We used PostgreSQL, a robust and secure relational database system, to store this data. Security measures were put in place to protect the data, including encryption at rest and in transit, regular backups, and strict access controls.

The Ingestor tool was then developed to convert the selected xAPI data points from the replicated LRS MongoDB into a relational table structure. The technology is comprising two major components: real-time data synchronization and backfilling. The real-time data synchronization component is written in Python. On the other hand, the backfilling component, responsible for filling historical data gaps, was developed using Golang due to its efficiency and performance advantages in handling large datasets. Together, these components make Ingestor a versatile tool for data conversion and synchronization.

Upon the successful establishment of the data pipeline, the commencement of the machine learning model training phase for the ECLAIR system was initiated. This intensive process was undertaken on high-performance computing systems, owing to the considerable amount of data involved. One representative example of the machine learning models employed within the ECLAIR system is the Bayesian Knowledge Tracing (BKT), a Python-based implementation of the BKT model. To ensure up-to-date integration of new learning log data, the model is refreshed at regular intervals, specifically every 30 minutes if new data has been detected by the system (Badrinath et al., 2021). Each model that is trained is then stored in the Recommender DB along with a timestamp showing when the model could first be used to give recommendations. As it is continuously being updated, historical version snapshots of the model are stored and can be used by researchers to replay or recreate recommendations of particular students at a point in time. This can be useful also if a research needs to investigate in detail why and how recommendations were generated previously by the model so that explanations can be checked and validated. It also allows developers and researchers to examine problems that have occurred in the past.

Finally, when a learner uses the recommendation panel on a satellite LEAF system, this sends a request via Kodama to the recommender module in the ECLAIR system.

---

\(^2\) [https://scikit-learn.org/](https://scikit-learn.org/)
Personalized recommendations and explanations are then generated based on the data and models stored in the central database, relayed back to the satellite LEAF system and displayed to the learner. The feedback mechanism ensured that the recommendations were continuously improved based on the learners’ interactions with the system. Also, as the recommendation model is centrally located, it enables researchers and developers to update and deploy new features and targeted interventions for evaluation with efficiency, instead of having to update multiple satellite LEAF systems.

In summary, the implementation and integration of the ECLAIR system into the LEAF infrastructure was a complex yet rewarding process. The successful integration has the potential to transform the learning experience by providing personalized, efficient, and engaging AI-driven recommendations. The modularity of the system also offers opportunities for future enhancements and scalability, as the need arises.

6. Conclusion and Future Work

The development, implementation, and integration of the ECLAIR system with the existing LEAF infrastructure has demonstrated the power of combining AI, machine learning, and data processing technologies to transform the learning experience. The architecture based on microservices provides an efficient and scalable solution, while the utilization of high-performance computing systems and advanced machine learning models like BKT facilitates the processing of vast amounts of data to generate personalized learning recommendations.

The implementation of ECLAIR has successfully tackled several challenges, including seamless data transmission, real-time data synchronization, and stringent security measures for data privacy. Furthermore, it has underscored the effectiveness of Python for developing AI applications, the efficiency of Flask framework in creating RESTful APIs, and the advantage of Golang in handling large datasets.

The use of the Ingestor tool for real-time data synchronization and backfilling of historical data, along with the frequent refreshing of the machine learning model, ensures that the learning recommendations are always based on the most current data. The feedback mechanism integrated into the LEAF system allows for continuous improvement in the recommendation quality, ensuring a more engaging and personalized learning experience.

Future work on the ECLAIR system could further extend its capabilities and scalability. Potential directions include enhancing the machine learning models to handle even more complex learning scenarios and incorporating additional forms of learning data, such as biometric or psychometric data, to refine the personalization of learning recommendations. There may also be potential for integration with other LMSs or learning platforms to broaden its impact.

In summary, the ECLAIR system, through its innovative use of AI and machine learning technologies, promises to be a transformative tool in the landscape of personalized learning. Its successful integration with the LEAF system demonstrates its potential, not just for improving the learner’s experience, but also for driving forward the field of AI-powered education as a whole. The flexible and scalable architecture of the system also ensures that it is well-suited for future enhancements and integrations, potentially creating an even greater impact on the education sector.

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

References


Sharing Learning Log while maintaining privacy over blockchain: Heuristic Evaluation of BOLL

Patrick OCHEJA a, Rwitajit MAJUMDAR a, Brendan FLANAGAN b, & Hiroaki OGATA a

aAcademic Centre for Computing and Media Studies, Kyoto University, Japan
bCenter for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University, Japan
*oceneja.ileanwa.65s@st.kyoto-u.ac.jp

Abstract: The Blockchain of Learning Logs (BOLL) system is a blockchain-based platform for connecting learners’ educational records from multiple schools. The BOLL system creates a permanent record of learners’ lifelong learning as immutable hashes on the blockchain, which can be analyzed to inform teaching and learning. This paper presents a usability analysis of the BOLL system using the 10 Jakob Nielsen Heuristics, with two user groups: students and teachers. The study evaluates the usability of various features, including the ability to view learner data from multiple schools, manage permissions, visualize analytics derived from connected learning logs, and provide access to learning materials used at various schools. Our findings highlight the successes of the BOLL system, including strong performance in areas such as consistency, real-world relevance, and user control. However, limitations were identified in error handling and the availability of comprehensive help and documentation. We conclude by emphasizing the need for future work to address these limitations and suggests potential avenues for improvement. Overall, this research contributes to the development of a user-friendly and privacy-conscious platform that can facilitate lifelong learning and enhance educational data sharing and analysis.

Keywords: BOLL, Blockchain, Learning Logs, Usability, Heuristic, Evaluation, Education

1. Introduction

The increase in the adoption of technology in teaching and learning has resulted in the need for a secure and efficient system to connect learners’ educational records across multiple schools. The Blockchain of Learning Logs (BOLL) system offers a promising solution based on blockchain technology to create a permanent and tamper-proof record of learners’ lifelong learning experiences (Ocheja, Flanagan, Ueda & Ogata, 2019). By maintaining privacy while enabling data sharing, BOLL has the potential to revolutionize the educational landscape. BOLL provides various features, interfaces and infrastructure to support access and usage of education data across different schools attended by learners including: learner profile, engagement analysis, cohort distribution, and access to learning contents as presented in (Ocheja, Flanagan, Majumdar, & Ogata, 2021; Ocheja, Flanagan, & Ogata, 2022a). The architecture and key components of the BOLL system distinguish it from other blockchain-based solutions. This decentralized platform utilizes a network of distributed nodes to validate and record learners’ educational activities. Consequently, BOLL ensures the integrity and security of the data, giving learners control over their educational records without compromising on trustworthiness and reliability. Understanding the usability of the BOLL system is crucial for its successful adoption and integration into educational environments. In light of this, we conducted a comprehensive usability analysis using the 10 Jakob Nielsen Heuristics (Nielsen, 1994). By evaluating the user experience from the perspectives of students and teachers, who engage with the system to access, share, and analyze educational records, we aim to identify strengths, weaknesses, and areas for improvement in the design and functionality of the BOLL system.
This research not only provides valuable insights into the usability of the BOLL system but also contributes to the broader field of blockchain-based educational systems. With a specific research focus on the usability of the BOLL system in maintaining privacy while securely sharing and analyzing educational records, we aim to address the research question of

1. What is the usability of the Blockchain of Learning Logs (BOLL) system?

2. Related work
The increase in the adoption of technology in teaching and learning has resulted in the need for a secure and efficient system to connect learners’ educational records across multiple schools. The Blockchain of Learning Logs (BOLL) system offers a promising solution based on blockchain technology to create a permanent and tamper-proof record of learners’ lifelong learning experiences (Ocheja, Flanagan, Ueda & Ogata, 2019). By maintaining privacy while enabling data sharing, BOLL has the potential to revolutionize the educational landscape. BOLL provides various features, interfaces and infrastructure to support access and usage of education data across different schools attended by learners including: learner profile, engagement analysis, cohort distribution, and access to learning contents as presented in (Ocheja, Flanagan, Majumdar, & Ogata, 2021; Ocheja, Flanagan, & Ogata, 2022a).

3. Methodology
The evaluation in this study is conducted through a heuristic approach, which involves experts systematically examining the BOLL system against the established usability heuristics (Nielsen, 1994). This approach allows us to identify potential usability issues and provide recommendations for enhancing the system's usability and user experience. In this evaluation, we define the following user goals to be carried out while evaluating the usability of the BOLL system.

**Students:**
- View learning data at different schools to support current learning activities.
- Grant and revoke permissions to view data at previous school to teacher at current school.
- Access digital contents such as textbooks and lecture slides as revision materials.

**Teachers:**
- View profile of students on prerequisite courses to support their current learning goal.
- View the engagement distribution of students in a prerequisite course towards recommending appropriate revision across cohorts.
- Compare the engagement of students across different times or courses to detect consistent behaviour or at-risk tendencies.
- Access digital contents used by students in a prerequisite course to understand the depth of learning or assessment.

3.1 Participants
A diverse group of participants was selected to represent the target user groups of the BOLL system: students and teachers. To minimize bias, participants were recruited from various educational institutions that could potentially use the BOLL system. Efforts were made to ensure a varied representation in terms of teaching and learning experience, as well as familiarity with blockchain technology and learning management systems. In total, 4 teachers and 7 students participated in this evaluation.

3.2 BOLL Features
To provide participants with a hands-on experience of the BOLL system, specific features and functionalities that support the user goals were demonstrated. These features included the ability to view learner data from multiple schools, manage permissions for data access, visualize analytics derived from connected learning logs, and provide access to learning
materials used across various educational institutions.
3.3 Questionnaire Design
For each user group, a set of questionnaires were developed based on the 10 Jakob Nielsen Heuristics for usability evaluation. These heuristics served as a set of guidelines for assessing the usability of interactive systems. Each heuristic was adapted to the context of the BOLL system, and specific questions were formulated to evaluate the system's performance against each criterion. Furthermore, the questionnaires aimed to gather both quantitative and qualitative feedback from participants regarding their experience and perception of the BOLL system's usability. To mitigate biases in this study, we developed clear and unbiased questionnaires with expert input and pilot-testing. Although the study had a limited sample size which affects generalizability, the results still offer valuable insights into the BOLL system's usability.

4. Results
The data collected from the questionnaires were analyzed to assess the usability of the BOLL system. Figures 9 and 10 show the results of the analysis of the quantitative data retrieved on a 5-scale Likert for each of the system features for each user group.

Figure 9. Results of the heuristic evaluation of 4 functionalities on BOLL designed for teachers.

To interpret the results of the heuristic evaluation conducted on a 5-point Likert scale, we first calculate the mean of the scores given by the participants for each of the 10 Jakob Nielsen heuristic criteria. For each mean score, we assign a meaning as follows:

- **4.5 – 5.0, Excellent:** The Boll system performs exceptionally well on this criterion.
- **3.5 – 4.49, Good:** The BOLL system generally meets the heuristic guidelines and achieves a satisfactory level of usability for this criterion.
- **2.5 – 3.49, Fair:** The BOLL system shows some usability issues and room for improvement in meeting the heuristic guidelines for this criterion.
- **1.5 – 2.49, Poor:** The BOLL system has significant usability issues and deviates considerably from the heuristic principles for this criterion.
- **<1.49, Very Poor:** The BOLL system severely lacks usability and fails to address the heuristic guidelines for this criterion.
We report the results of the evaluations of various features of BOLL by each user group below.

4.1 Teacher group
Overall, the heuristic evaluation conducted by the teacher group revealed a generally positive perception of the BOLL system in terms of usability. The learner profile view received positive ratings across most of the evaluated heuristics, indicating a good level of consistency, error prevention, recognition over recall, and match to the real-world context. The visibility, user control & freedom, flexibility & efficiency, and aesthetic and minimalist aspects of the interface were well-regarded, receiving ratings of 4.0 or higher. However, there is room for improvement in terms of help & documentation, which was rated as fair (m = 2.50). The evaluation of the past engagement view, compare engagement view, and learning contents view also showed generally positive results, with consistent ratings above 3.0 for most heuristics. Notably, the flexibility & efficiency heuristic scored high in the compare engagement view (m = 4.50), indicating its effectiveness. Overall, these findings suggest that the BOLL system demonstrates promising usability, with specific areas identified for potential enhancements, particularly in terms of providing better help and documentation resources for teachers.

4.2 Student group
The heuristic evaluation conducted with the student group demonstrated overall positive perceptions of the usability of the BOLL system. In the student’s data view, the system received high ratings for consistency and standards (m = 4.29), match to the real-world context (m = 4.57), visibility (m = 4.00), and aesthetic and minimalist design (m = 4.14), indicating that students found these aspects well-designed and effective. The system also performed well in terms of user control and freedom (m = 3.71), recognition over recall (m = 3.71), and flexibility and efficiency (m = 4.00). However, there were areas identified for improvement, particularly in recognizing, diagnosing, and recovering from errors (m = 2.86) and providing comprehensive help and documentation (m = 3.14). Similarly, in the data permission view, the system exhibited strengths in consistency and standards (m = 3.43), user control and freedom (m = 3.71), visibility (m = 3.86), and aesthetic and minimalist design (m = 4.00). However,
there were opportunities for improvement in recognizing, diagnosing, and recovering from errors (m = 2.86) and providing comprehensive help and documentation (m = 2.57). In the learning contents view, the system received positive ratings for consistency and standards (m = 3.43), user control and freedom (m = 3.71), error prevention (m = 3.71), recognition over recall (m = 4.14), visibility (m = 4.71), and aesthetic and minimalist design (m = 4.00). However, improvements were needed in recognizing, diagnosing, and recovering from errors (m = 2.57) and providing comprehensive help and documentation (m = 2.86). Overall, the evaluation provided valuable insights, indicating that while the BOLL system demonstrated strengths in various usability aspects, there were specific areas, such as error handling and help/documentation, that required attention and improvement. These findings highlight the importance of addressing these concerns to enhance the overall user experience for students.

5. Conclusion

In conclusion, this paper presented a heuristic evaluation of the Blockchain of Learning Logs (BOLL) system, focusing on its usability from the perspectives of both students and teachers. The evaluation revealed several successes of the BOLL system, including strong performance in areas such as consistency, real-world relevance, user control, and aesthetic design. These findings demonstrate the system’s potential to effectively support lifelong learning and facilitate the sharing of educational records securely. However, limitations were identified, particularly in terms of error handling and the availability of comprehensive help and documentation resources. Future work should prioritize addressing these limitations to further enhance the system’s usability. Additionally, the small sample size and potential biases in participant selection limits the generalizability of the findings in this study. However, the insights gained from this study, in conjunction with previous research and theoretical considerations, provide a valuable foundation for further explorations and refinements of the BOLL system’s usability. Future studies with larger and more diverse samples would help validate and extend our findings.

Acknowledgements

This work was supported in part by the following grants: Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Fellows JP22KJ1914, (B) JP20H01722, JP23H01001, (Exploratory) JP21K19824, (B) JP22H03902 and New Energy and Industrial Technology Development Organization (NEDO) under Grant JPNP20006 and Grant JPNP18013.

References

Program Visualization System Supporting Teacher-Intended Stepwise Refinement

Koichi YAMASHITA*, Hiroki SOMA, Satoru KOGUREb, Yasuhiro NOGUCHIb, Raiya YAMAMOTOa, Tatsuhiro KONISHib & Yukihiro ITOHc

aFaculty of Business Administration, Tokoha University, Japan
bFaculty of Informatics, Shizuoka University, Japan
cShizuoka University, Japan

*yamasita@hm.tokoha-u.ac.jp

Abstract: This study portrays a learning support system that visually represents the structure of stepwise algorithm refinement, also known as the structure of program code abstraction, aligned with the teacher's instructional objectives. We focus on the varying needs of learners in programming education and consider the possibility that different levels of stepwise refinement may be required based on their knowledge and experience. Novice learners with limited programming experience may need program code abstractions in finer steps, rather than the abstraction of numerous statements simultaneously. For learners with a certain level of experience, abstraction with excessive steps may hinder their holistic understanding of the entire program from a higher perspective. In this study, we developed a program visualization system that empowers teachers to customize the visualizations of the structure of stepwise abstractions and the data structures utilized by the target program according to their instructional intentions. We introduced the proposed system into an actual class and evaluated its effectiveness by measuring the degree of improvement in learners' understanding based on tests.

Keywords: Programming education, program visualization system, stepwise refinement and abstraction, educational authoring tool

1. Introduction

To date, various program visualization (PV) systems have been developed to support the programming learning of novice learners (Sorva, Karavirta, & Malmi, 2013). We had developed a PV system called Teacher’s Explaining Design Visualization Tool (TEDViT) and conducted several classroom sessions introducing this system into actual classes (Kogure et al., 2014). There are two features of TEDViT that distinguish it from other existing systems: first, it enables teachers to customize PVs based on their objectives, and second, it enables learners to observe and compare two PVs with different abstraction levels of target data structures. However, it cannot provide visualizations at different levels of abstraction for program-code that represents concrete operational procedures for the target domain world of the data structures to be processed.

In programming education, stepwise refinement has been used to design program codes by gradually decreasing the level of abstraction from highly abstract algorithms. Moreover, there have been various attempts to support learners’ understanding of programs by focusing on stepwise abstraction used to understand the algorithm behind program code by gradually increasing the level of abstraction of the code. In a classic study, Shneiderman & Mayer (1979) summarized the procedures that expert programmers undertake during code reading. These steps include chunking some statements within the target program, recognizing the function of the chunk, and progressively assembling the chunks to form larger units. The chunking process is continued recursively until the entire program is understood. This implies that learners need to perform stepwise abstraction by chunking concrete operation procedures to understand a program. However, the literature suggests
that novice learners may frequently struggle to comprehend a program owing to difficulties in effectively chunking the code.

In this study, we developed a system that can visualize program behavior by drawing target program codes that represent concrete processing procedures and problem analysis diagrams (PADs) that represent algorithms with a higher level of abstraction than program code, enabling users to gradually change abstraction level of PADs. The abstraction level of PAD here refers to the size of blocks of program code to be chunked, with larger chunked blocks being considered highly abstract and smaller chunked blocks being considered less abstract. This system is an extension of TEDViT that enables teachers to define the chunks to be included in algorithm visualization. Learners understand chunks differently depending on their knowledge and experience. Hence, our system, which allows teachers to define chunking structures, is expected to provide PVs that better align with learners' understanding compared to existing systems with fixed chunking structures determined by developers.

2. Related Works

2.1 TEDViT

Price, Baecker, & Small (1998) categorized users of PV systems into the following four roles: User/viewer who observes the world visualized by the system; visualizer/ animator who defines the visualization of the target program's behavior; software developer who develops the PV system; and programmer who designs the program that is the target of visualization. A typical existing PV system assigns the role of user/viewer to the learners, visualizer/animator and software developer to the PV system's developer, and programmer to the learners and teachers.

As mentioned earlier, we had developed a PV system called TEDViT, which assigns the role of visualizer/animator to the teachers. TEDViT is a system that enables teachers to define drawing rules to customize the visualization suitable for each target program and for the target user/viewer. When a teacher explains a program behavior to novice learners using visualizations, the teacher may sort out, highlight, and lay out some drawing objects from the target data structure. The teacher may also attach descriptive natural language texts to some objects to assist students' understanding and may draw pointers and array indexers using arrows. We consider these visualizations to be based on a policy that the teacher determines according to the learners' background knowledge and degree of understanding, the properties of the target program, etc. We refer to these visualization policies and the concepts that the teacher intends to explain using these policies as the teacher's "intent of instruction." TEDViT is, therefore, a PV system that can reflect teachers' intent of instruction regarding visualizations by assigning them the role of visualizer/animator.

TEDViT enables visualizers/animators to create or edit a configuration file for PV creation based on their intent of instruction independently from the target program file. Configuration file comprises a set of drawing rules, each of which is a comma-separated value (CSV) entry consisting of a condition part and an object part. The condition part defines the condition to fire the drawing rule. The object part defines the operation to edit the target object (i.e., "create," "delete," and "update" are available) and the attributes necessary to draw the object such as the object type, position, color, and corresponding variables. TEDViT generates PVs by scanning configuration files, interpreting the visualization policy represented by a set of drawing rules, and visualizing drawing objects accordingly. Users/viewers can observe the program behavior based on the PVs. The supporting programming language is C.

2.2 Programming Learning Support Based on Stepwise Refinement and Abstraction

In general, a PV system visualizes the target domain world involved in program processing and aims to support learners in understanding the program's meaning by enabling them to observe the changes in the domain world caused by the program. Each statement is used as
a unit of meaning, that is, a unit that causes changes in the domain world. Although PV systems can visually support the understanding of the meaning of each statement, they may not be able to provide sufficient support for abstracting the meaning of multiple statements and understanding it, and hence, understanding the meaning of the entire program.

The stepwise refinement technique has been widely used in programming education. The procedure summarized by Shneiderman & Mayer (1979) implies that program comprehension can be achieved by the opposite operation, namely, stepwise abstraction from the detailed meaning of individual statements to abstract meaning that spans multiple statements. Here, abstraction is the interpretation of the meaning of chunking and chunks of the program, and refinement is the decomposition of chunks into smaller chunks. In learning programming, it is also necessary to understand the meaning of programs through the abstraction of operational procedures and refinement of algorithms; this is not supported by PV systems. Based on this idea, there have been several attempts to aid novice programming learners by supporting the stepwise abstraction of programs and pseudocode (Shinkai & Sumitani, 2007; Watanabe, Tomoto, & Akakura, 2015; Kogure et al., 2013).

3. Proposed System

In this study, we focused on the possibility that the learners may need different granularities of chunking in programming learning based on stepwise refinement depending on their knowledge and experience. For novice learners with minimal programming experience, small chunks of program code have to be abstracted at multiple levels of abstraction, rather than chunking a large amount of code simultaneously. For learners with a certain level of experience, chunking extremely granularly may hinder their holistic understanding of the entire program from a higher perspective. Yamashita et al. (2016) highlighted that there are two possible types of abstraction of sequences of operations: simple grouping and process generalization. In our teaching experience, learners with insufficient knowledge of CS and mathematics often struggle to understand chunking, particularly when it involves process generalization. In such cases, teachers could customize the chunking granularity and explanation of the meaning of the chunks based on the differences in learners' knowledge and background. Moreover, the differences in the chunking granularities would have a certain effect on their visualization policy of the target domain world.

In this study, we developed a PV system that enables teachers to freely define and visualize the chunks of program code and their structures based on their intent of instruction. The chunks of program code are visualized as PAD-based algorithmic representations. Our system can enable users to increase the abstraction granularity of each chunk object and visualize more detailed PADS. It builds upon TEDViT, a program visualization (PV) system that empowers teachers to customize the visual representations of the domain world according to their instructional objectives. Therefore, our system provides teachers with the flexibility to define and customize the visualizations of the data structures involved in program processing. Figure 1 provides a screenshot of the proposed system.

Our system generates a learning environment with four main visualization areas. In area (A), the target program code is visualized, and an executing statement or chunk is highlighted. In area (B), a PAD-based algorithm representation is visualized. Area (C) visualizes the currently executed operation, displaying the text of the executing chunk. Area (D) visualizes the domain world, drawing the data structure of the program's processing target based on the definition file edited by the teacher as in TEDViT.

Each element of the PAD in area (B) corresponds to an abstracted chunk of statements in a certain range of the program code. The correspondence between each element of the PAD and each range of the program code can be identified by the number assigned to each element of the PAD and each range of the program code. In addition, buttons such as "+" and "-" are provided at the top of each element of the PAD. By clicking these buttons, the chunk of each element can be displayed in one more step in detail, as shown in Figure 2. As the algorithm representation visualized in area (B) includes chunking structures, it cannot strictly be called a PAD. We call the algorithm representation visualized...
by the proposed system "extended PAD." In addition to the aforementioned features, some elements of the extended PAD have buttons for "step in function" and "step over function," resembling typical debuggers. Some elements also have buttons for "execute all loop" and "execute one round of loop" depending on the processing contents.

Figure 1. Screenshot of the proposed system.

Figure 2. Visualizing a chunking structure in one step more detailed.

4. Evaluation

To assess the effectiveness of our system, we conducted a classroom practice by incorporating it into an actual programming course. The hypothesis to be verified is whether
our system could help learners to understand the behavior of programs and algorithms in code reading learning better than an existing PV system, TEDViT. The actual class whereby the systems were introduced was an on-demand class offered to second-year university students majoring in computer science. The number of participants in our class was 47, and they had 1-2 years of programming experience. Two types of sample programs, quick sort and merge sort, were prepared for code reading. All participants had already learned quick sort in the previous semester but had not learned merge sort.

In the practiced class, we had the participants perform code reading learning in the order of quick sort first and merge sort second. In each code reading learning, they took a pre-test before learning with PV systems, and then performed two learnings of code readings, one with TEDViT and the other with our system. They took tests immediately after each of the two learnings, which were positioned as a mid-test and a post-test, respectively. To reduce order effects, we divided the participants into two groups, Groups A and B, with each group learning with PV systems in reversed order. We also reversed the order of use of the system in the two learnings of quick sort and merge sort for each group as well. After all the learnings were conducted, we conducted a questionnaire survey.

In the code reading learning using TEDViT, participants observed changes in the logical data structure visualized in the domain world with stepwise execution of each statement of the target program. In the learning using our system, participants used the ‘+’ and ‘−’ buttons provided in area (B) of Figure 1 to observe the chunking structure of the extended PAD and simultaneously observed the changes in the domain world with stepwise execution of each statement or each chunk. The pre-test, mid-test, and post-test consisted of nine questions for the quick sort and 11 questions for the merge sort and were identical for each target program. Participants were required to take six tests in the entire class, and a total of 32 participants (13 in Group A and 19 in Group B) took all six tests.

Table 1 summarizes the average correct answer rates for each test for the 32 participants who took all tests. Underlined scores are the rates for the tests immediately after learning with our system. Both Groups A and B, regardless of the target program, demonstrated improved scores after code reading learning compared to before learning, suggesting that participants would improve their understanding of the programs. We also calculated average improvements for learning with TEDViT and for learning with our system by calculating score differences for all test results compared to the results of the immediately preceding test. We found the average improvement of correct answer rates for learning with TEDViT was 9.2% and that for learning with our system was 14.4%. Currently, we have not obtained any statistical results that significantly support our hypothesis that our system contributes more to learners’ understanding of programs than TEDViT. However, the observed average improvement of correct answer rates for our system, which is 5% higher than that of TEDViT, suggests a significant degree of effectiveness in enhancing learners’ understanding.

Table 1. Average correct answer rate of pre-, mid-, and post-test

<table>
<thead>
<tr>
<th>Group</th>
<th>N.</th>
<th>Target</th>
<th>Pre-test</th>
<th>Mid-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13</td>
<td>Quick sort</td>
<td>.554</td>
<td>.662</td>
<td>.708</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Merge sort</td>
<td>.608</td>
<td>.746</td>
<td>.738</td>
</tr>
<tr>
<td>B</td>
<td>19</td>
<td>Quick sort</td>
<td>.332</td>
<td>.668</td>
<td>.679</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Merge sort</td>
<td>.526</td>
<td>.768</td>
<td>.758</td>
</tr>
</tbody>
</table>

The post-class questionnaire included a 5-point pairwise comparison of whether TEDViT or our system was easier to learn, with 1 indicating a strong preference for TEDViT, 5 indicating a strong preference for our system, and the median being 3. The mean value for this item was 3.3 for the 30 participants who responded to the questionnaire, indicating that participants tended to prefer our system to TEDViT. Moreover, the mean values of all the items for which participants were asked to subjectively evaluate the effectiveness of the functions of our system were also higher than the median. Subjective evaluations of the
learning effectiveness of our system were also favorable, suggesting that participants positively viewed our system.

5. Conclusion

This study described a PV system that can simultaneously provide learners with processing procedures and processing objects at different levels of abstraction. The system achieves this by generating visualizations of the stepwise chunking structure of program code and visualizations of logical data structure, which are the objects being processed. We assessed the effectiveness of our system by introducing it in an actual class. As our system is based on the existing PV system, TEDViT, we asked participants to perform code reading learning using TEDViT and learning using our system and measured the effects of two learning activities by administering pre-, mid-, and post-tests. To reduce order effects, participants in the practice class were divided into two groups, each with an opposite learning sequence. In addition, two target programs for code reading, quick sort, and merge sort, were prepared, and the effect of the systems was measured on multiple program codes. The evaluation results based on the correct answer rates in three tests indicated no statistical difference between code reading using TEDViT and code reading using our system. However, TEDViT achieved a 9.2% improvement in correct answer rate on average, while our system achieved a 14.4% improvement in correct answer rate on average, indicating that the improvement was more than 5% better for learning with our system. We can conclude that the results indicate a certain degree of support for the effectiveness of our system.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Numbers JP19K12259 and JP22K12290.

References


Automatic Detection of Negotiation in Collaborative Complex Problem Solving Interactions

Daevesh SINGH*, Ulfa KHWAJAA, Sahana MurthyA, & Ramkumar RAJENDRAN

AIDP in Educational Technology, Indian Institute of Technology Bombay, INDIA
*daeveshsingh@iitb.ac.in

Abstract: When learners collaborate on complex problems and open-ended tasks, the mechanism of negotiation plays a crucial role in establishing a common understanding and achieving a shared goal among them. Research has shown that negotiation improves problem-solving processes, making it an essential skill to be developed among learners. In this study, we propose a method for automating the identification of negotiation in learners’ discourse during collaboration. We leverage language models like BERT, RoBERTa, and GPT2 along with traditional machine learning models like logistic regression to detect utterances of negotiation in learners’ discourse while they collaboratively solve engineering estimation problem in an Open-Ended Learning Environment (OELE) called Modeling Based Estimation Learning Environment (MEttLE). Our findings suggest that our approach can accurately identify negotiation utterances with a high accuracy of 0.924 and 0.781 kappa value with a relatively smaller training set. Our method is the first step in real-time detection of negotiation, thereby enabling educators to design scaffolds and environments to help learners engage in effective negotiations.

Keywords: negotiation, open-ended learning environments, collaboration, complex problem-solving.

1. Introduction

Collaborative problem-solving (CPS) is a vital skill involving multiple individuals working together to devise and sustain collective solutions to challenges (Fiore et al., 2017). CPS is crucial as the modern workforce increasingly demands seamless collaboration within diverse teams, efficient exchange of expertise, and effective communication across disciplinary boundaries (Pugh, Rao, Stewart, & D'Mello, 2022). To foster this skill, it is imperative to create opportunities for students to engage in productive CPS activities. Despite the positive impact of CPS on both academic and social educational outcomes, significant challenges persist when implementing CPS effectively within everyday classroom settings (Johnson, Johnson, & Smith, 2007). Teachers encounter difficulties in structuring group interactions and monitoring fruitful collaboration (Van Leeuwen, Janssen, Erkens, & Brekelmans, 2013). Students also face hurdles in CPS, including uneven participation in group tasks (Freeman & Greenacre, 2010) and deficiencies in communication and collaborative skills (Li & Campbell, 2008; Pauli, Mohiyeddini, Bray, Michie, & Street, 2008).

To address these challenges, it is imperative to pinpoint and assess essential CPS processes exhibited by learners and offer diagnostic feedback (Fiore et al., 2017). This paper focuses on automating the identification of a critical CPS process—negotiation—while students collaboratively tackle complex problem within an Online Environment for Learning and Education (OELE) called Modeling Based Estimation Learning Environment (MEttLE). This automated analysis will subsequently empower educators to provide valuable feedback to learners.

The mechanism of negotiation is vital for effective collaboration, particularly in complex problem-solving domains, enabling collaborators to employ diverse strategies and
reach agreements through critical evaluation, justification, sensemaking, and co-construction of solutions (Hesse et al., 2015). Previous research has demonstrated the productive behaviours exhibited by learners when engaging in negotiation within OELEs during complex problem-solving tasks (Khwaja & Murthy, 2022). Therefore, fostering negotiation in such environments is crucial for supporting learners’ problem-solving processes. However, identifying negotiation utterances in learners’ discourse is challenging, particularly when the interactions are complex and multi-layered. Also, the manual identification of negotiation utterances is time-consuming and impractical for large-scale studies.

To address this gap, researchers have proposed a method for automating the identification of negotiation and other collaborative mechanisms in learners’ discourse during the process of collaborative problem-solving (Flor et al., 2016; Pugh, Rao, Stewart, and D'Mello, 2022; Hao et al., 2017). We extend this work by leveraging state-of-the-art language models like BERT, RoBERTa, and GPT2 to automatically detect negotiation utterances in learners’ discourse while they solve open-ended problems. Results show that GPT2 emerged as the best-performing model with an accuracy of 0.924 and a kappa value of 0.781. The novelty of our research lies in the fact that the models employed here demonstrate improved performance compared to existing approaches in automating the identification of negotiation utterances in learners’ discourse during collaborative problem-solving. Additionally, the study presents a unique context by focusing on the engineering estimation problem, which is a frequent practice used by engineers and scientists when exact data and precise governing equations are unavailable (Mahajan, 2014).

**Figure 1.** The figure shows a collaboration between two students, and the discourse data is captured and converted to transcripts. This transcript is then analysed by the ML models for generating discourse analytics for teachers and personalised feedback for students, which is first delivered to the teacher, who has the agency to modify it before delivering it to students. The study constitutes a part of a larger research that aims to enhance productive behaviours in collaborations by providing students with personalised feedback (Fig. 1). Here, the teacher will have access to the discourse analytics affording them the agency to modify the feedback provided by the models. The overarching goal is to attain insightful perspectives on the evolution of different and essential collaborative learning mechanisms during collaboration. This can help educators design scaffolds to enable learners to engage in productive behaviours.

The paper is structured as follows: Section 2 presents the background and literature synthesis, followed by the OELE description in Section 3. We then delineate the research goal and procedure in Section 4. Subsequently, we present the results in Section 5, leading to the discussion and conclusion in Section 6.
2. Background and Related Work

In this section, we first highlight the significance of negotiations in the process of collaborative problem-solving. The second sub-section discusses the theoretical CPS framework that we have employed for the manual coding of the data. Then, we discuss the state-of-the-art text classification algorithms. Finally, we discuss different studies that aim to detect negotiation in discourse data.

2.1 Negotiation and Problem-solving

Effective collaboration in Computer Supported Collaborative Learning (CSCL) research necessitates negotiation when dealing with divergent opinions and ideas from peers. Negotiation plays a crucial role in complex problems and open domains that require critical examination, the convergence of ideas, and contributions from all collaborators to arrive at a suitable solution (Carell & Herrmann, 2009). In this context, it can be defined in various ways, including resolving conflicts, attempting to agree on goals, and critically examining different perspectives (Baker, 1994; Fleck et al., 2009). Learners engaged in a negotiation must present their ideas, defend their positions, and use various strategies to reach an agreement. This process must ultimately result in a shared understanding of the problem to ensure that all collaborators have reasoned and agreed upon a common ground in problem-solving (Beers, 2006). Through negotiation, learners can engage in meaningful discussions and sensemaking, exchange ideas, learn from their peers, and contribute to productive outcomes. It is particularly vital in OELEs, where learners need to explore new ideas, tools, and critically examine complex problems. Therefore, it is essential to identify and understand negotiation utterances and design scaffolds to foster it among learners to support their problem-solving abilities.

2.2 Collaboration Learning Mechanisms Framework to identify negotiation

To identify and comprehensively understand instances of negotiation among learners within MEttLE (Kothiyal & Murthy, 2018), we employed the Collaboration Learning Mechanisms (CLM) framework developed by Fleck et al. (2009). This framework centres on two pivotal mechanisms for collaborative learning within open problem contexts: coordination and discussion. Within these mechanisms, the CLM framework emphasises how specific actions and verbal interactions represent different aspects of collaboration and how they collectively contribute to productive outcomes in collaborative problem-solving activities. The collaborative discussion mechanism entails the exchange of information, ideas, and negotiation of meaning, while the coordination mechanism underscores joint attention, awareness, and narration. This framework helped us identify and understand the mechanism of negotiation that involved verbal discourse of disagreeing, conflicting, proposing alternatives, criticising, explaining, and justifying one’s idea, and actions such as undoing, deleting, nodding head to depict disagreement, etc. CLM defines negotiation as a conversation that involves all parties being critically and constructively engaged with each other’s proposals for joint consideration. We then further added 9 more mechanisms of collaboration based on our data. To ensure the reliability and validity of our approach, two researchers independently calculated the inter-rater reliability for 10% of the dataset, resulting in a robust kappa score of 0.7, indicative of strong agreement between the researchers. Among the various collaborative mechanisms, we observed that learners were engaged in negotiation approximately 31.5% of the time during their problem-solving activities.

Table 1. The table presents the episode of Negotiation from Group 2, where both the learners S3 and S4 read the problem statement and immediately get into a discussion on how to solve it. S3 is seen to propose a solution method that S4 does not easily accept.

<table>
<thead>
<tr>
<th>Student</th>
<th>Utterences</th>
<th>Time</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

443
2.3 Text Classification

Text classification, a machine-learning technique used to categorise open-ended text into predetermined categories, has gained significant attention in research. Its application to coding discourse data addresses the time and labour-intensive nature of manual coding processes. While manual coding can consume days of effort, machine learning models can accomplish the same task within a few minutes, thereby substantially reducing the required time and resources.

Recent advancements in large language models have presented opportunities to enhance classification accuracy further. Among these models, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) have emerged as a notable text representation model specifically designed for natural language processing. BERT distinguishes itself from previous models by generating dynamic and contextualised word representations through unique training tasks, such as masked-language modelling and next-sentence prediction. Leveraging the mechanism of transfer learning, BERT can be readily adapted to downstream tasks with relative ease.

In addition to BERT, another influential language model is RoBERTa (Liu et al., 2019). RoBERTa, based on transformer architecture, has garnered attention for its remarkable language generation capabilities. Pretrained on a massive corpus of text data, RoBERTa exhibits exceptional performance across various natural language processing tasks, including text classification. By capturing contextual dependencies and semantic relationships between words and phrases, RoBERTa effectively comprehends and generates coherent text.

Another prominent language model worth mentioning is GPT2 (Generative Pretrained Transformer 2). GPT2 is a transformer-based model known for its impressive language generation capabilities. It is trained on a massive corpus of text data and has demonstrated exceptional performance in a wide range of natural language processing tasks, including text classification. GPT2 captures the contextual dependencies and semantic relationships between words and phrases, enabling it to understand and generate coherent text.

To summarise, the availability of powerful language models such as BERT, GPT2, and RoBERTa offers researchers the opportunity to enhance the accuracy of text classification. Their proficiency in capturing contextual information, coupled with their transfer learning capabilities and demonstrated success across various natural language processing tasks, positions them as promising options for automating the categorisation of discourse data.

2.4 Collaboration Analytics
Collaboration Analytics pertains to the methodologies and strategies employed in the automated or semi-automated capture, analysis, mining, and extraction of data concerning interactions among collaborators (Schneider et al., 2021). We ground our work in the existing domain of linguistic modelling of collaborative problem-solving. In this sub-section, we shall discuss the existing research that makes use of advanced NLP methods to analyse discourse data in CPS. This discourse data emerge either from the text chats or is transcribed from the speech) and is used to model various skills during CPS, including negotiation, regulation, argumentation etc.

The paper by Flor et al. (2016) presents a study that investigates the use of automated techniques for accurately categorising interactions in collaborative problem-solving within simulated science tasks. It applies natural language processing (NLP) methods to analyse and classify these interactions using Naive Bayes (NB) and Hidden Markov Models (HMM). The results reveal the effectiveness of the automated classification models in accurately identifying and categorising various types of interactions, achieving an average accuracy of 59.2%.

Pugh, Rao, Stewart, and D'Mello (2022) investigated the feasibility of detecting seven cognitive and social CPS skills (including sharing information, negotiation, etc.) in classroom and lab settings. The participants were middle school kids who collaboratively solved math and physics problems. They achieved an AUROC value of 0.78 in the classroom setting and 0.83 in the lab setting using BERT.

The work by Hao et al. (2017) focuses on developing an automated annotation system called CPS-rater. This tool annotates the discourse into various CPS skills like negotiation, regulation etc. They have used different models, including Random Forest (RF) and NB. Linear chain conditional random field emerged as the best model, achieving an accuracy of 0.732 and a kappa value of 0.636.

To summarise, several studies have contributed to the field of automated analysis and classification of interactions in collaborative problem-solving (CPS) activities. Building upon these previous works, our study aims to extend the existing research by introducing novel contributions. Firstly, we focus on the specific context of the engineering estimation problem within collaborative problem-solving. This contextual specificity provides unique insights and findings applicable to the domain of engineering problem-solving. Secondly, our work employs advanced machine learning models that have demonstrated improved performance compared to existing approaches. By utilising these models, we aim to enhance the accuracy and efficiency of automated identification and categorisation of negotiation utterances in learners' discourse during collaborative problem-solving. These advancements contribute to the broader understanding and advancement of automated analysis techniques in the domain of CPS, addressing the need for more precise and effective analysis of collaborative interactions.

3. Learning Environment

The Modelling-Based Estimation Learning Environment (MEttLE), as shown in Figure 2, is an open-ended learning environment designed to scaffold novice learners in their estimation problem-solving. The tool's five sub-goals trigger a model-based estimation process that is essential for solving the estimation problem (Figure 2a). The three-phased model-building sub-goals, which include functional, qualitative, and quantitative aspects, help learners create, contextualise, and evaluate models of complex problems, one calculation, and one evaluation. MEttLE also includes metacognitive prompts (Figure 2b) that encourage learners to reflect on their models and problem-solving processes, as well as simulators (Figure 2c), hints, an info centre, guide me, question prompts, and a problem map to facilitate the modelling process. Novice learners have the flexibility to choose any path and revisit any sub-goals at any time in MEttLE.
To solve the estimation problem in this study, the solver is tasked with estimating the power of a car while comprehending the underlying problem system. To achieve this, they must analyse how the car behaves under the given operating conditions, determine the dominant parameters for these conditions, and create an equation involving the dominant parameters. Additionally, they must make assumptions and approximations to create and determine a simplified model of the problem context and its corresponding mathematical equation. One noteworthy aspect of METTLE is its open design, which provides immense opportunities for collaborative learning wherein learners undertake meaningful discussions and negotiations to solve this problem.

4. Methods

4.1 Participants

The participants in this study were three dyads comprising two female and four male students (age range of 18-21 years) from 3rd or 4th-year undergraduate engineering colleges in Mumbai. Students from mechanical or electronics engineering were chosen as participants, as the problem in METTLE required knowledge specific to these disciplines. The students were paired randomly based on their availability based on a pre-study Google form that collected their demographic details. IRB approval from our institution was sought.

4.2 Procedure and Data Collection and Data Analysis

The study was conducted in a research lab setting where the three engineering undergraduate dyads participated on different days. The researcher explained the study goal, the procedure, and the data that would be collected, after which the participants signed the consent form. The lab setup involved a single computer with METTLE for both participants, who had to solve the estimation problem in it collaboratively. They were asked to articulate their views explicitly during the collaboration. On average, the time taken by pairs to interact with METTLE was 90 minutes. Data was collected from multiple sources, like screen recording using OBS Studio (OBS, 2021), which ran on a laptop containing METTLE as well as audio and video data; however, for this study, we only used the transcription of their audio data. Learners’ discourse data was then passed to an AI tool called Otter.ai for speech-to-text transcription. Further, researchers coded the data using the CLM framework to identify learners’ negotiation instances.

4.3 Feature extraction, model training, and evaluation
In this study, we employed three traditional machine learning classifiers, namely Logistic Regression (LR), NB, and RF. We used the features obtained via TF-IDF (Spärck Jones, 1972) and the Word2Vec (Mikolov et al., 2013) approach. The intention of using these classifiers was that they are employed in most of the research in the domain of learning analytics. Additionally, we used three state-of-the-art pre-trained neural networks for natural language processing that include BERT, RoBERTa, and GPT2. These models were trained for binary classification problem, i.e. to classify discourse data into two categories: the presence of negotiation (labelled as 1) and the absence of negotiation (labelled as 0). We shall first discuss the feature extraction, model training of traditional machine learning models before delving into large language models.

TF-IDF Feature Extraction:

We employed the TfidfVectorizer from the sklearn.feature_extraction.text module to convert the text data into numerical features. The vectoriser was configured with parameters such as a minimum document frequency of 5 and a maximum of 10,000 features. This ensured that only frequently occurring words were considered, and the feature dimensionality was controlled.

In addition to TF-IDF, we trained a Word2Vec model using the gensim library. The Word2Vec model learned word embeddings from the training sentences with specific hyperparameters, such as size = 100, window = 5, min_count = 1, and workers = 4. These parameters were selected based on prior experimentation and domain knowledge.

The training data was preprocessed and divided into two sets: one for the TF-IDF-based approach and the other for the Word2Vec-based approach. For TF-IDF, the training data was transformed into TF-IDF features using the fitted vectoriser. For Word2Vec, the training data sentences were tokenised and converted into word embeddings using the trained Word2Vec model. The target variable was created by encoding the labels as integers.

We trained three classifiers, LR, RF, and NB, on the training data. We evaluated the performance of the classifiers on the test data using various metrics like classification accuracy, F1-score, and kappa value (explained at the end of this section). For BERT, we tokenised the texts using the Hugging Face Transformers library's tokeniser function and converted them to input features with padding and truncation. We split the data into a training set and a validation set and labelled each instance of discourse as either negotiation or non-negotiation. We then created training and validation datasets as PyTorch TensorDatasets, using the input features and labels. We fine-tuned the pre-trained BERT model for our specific task of classifying negotiation and non-negotiation discourse data. We defined the training parameters, such as batch size, number of epochs, and learning rate, and used the AdamW optimiser and the Cross-Entropy Loss function. In our case, we considered a batch size of 16, and the number of epochs was taken to be four. The learning rate was set as 2*(10)-5.

We then created data loaders for the training and validation datasets and trained the model on the training set for a fixed number of epochs using mini-batches. During each epoch, we optimised the model's weights using backpropagation and stochastic gradient descent.

To evaluate the performance of the model, we used the validation set, which the model had not seen during training. The same procedure was applied for RoBERTa and GPT2 as well, and the hyperparameters were also chosen to be the same. For performance evaluation, we chose the following four metrics:

1. Classification accuracy (CA): the ratio of the number of correct classifications to the total number of classifications.
2. F score (F1): is the harmonic mean of recall and precision.
3. Kappa (κ): measures agreement between the actual and the predicted labels by considering the by-chance prediction. κ is calculated using Eq. (1), where P0 is the overall accuracy of the model and Pe is the measure of the agreement between the model predictions and the actual class values as if happening by chance.

\[
\kappa = \frac{P_0 - P_e}{1 - P_0} \quad (1)
\]
5. Result

This section describes the results of the prediction models developed for classification.

Table 2. The evaluation metric of performance of ML models on “negotiation” classification.

<table>
<thead>
<tr>
<th>Text Features</th>
<th>Classifier</th>
<th>CA</th>
<th>F1</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>LR</td>
<td>0.818</td>
<td>0.777</td>
<td>0.283</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.794</td>
<td>0.712</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.903</td>
<td>0.893</td>
<td>0.664</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>LR</td>
<td>0.784</td>
<td>0.6890</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.369</td>
<td>0.355</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.873</td>
<td>0.853</td>
<td>0.530</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>0.907</td>
<td>0.906</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td>RoBERTa</td>
<td>0.889</td>
<td>0.889</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>GPT2</td>
<td>0.924</td>
<td>0.924</td>
<td>0.781</td>
</tr>
</tbody>
</table>

The performance of various machine learning and deep learning models was evaluated on the task of classifying discourse data into negotiation and non-negotiation. The models were trained on a dataset with 1989 instances and tested on a dataset with 686 instances, and this data split was done on a group level, i.e., data from two groups was used for training, and it was tested on the discourse of the remaining group. The dataset was preprocessed using standard NLP techniques like tokenisation, stop-word removal, and stemming. Two types of features were used for training the traditional machine learning models: TF-IDF and Word2Vec.

The machine learning models used in this study were LR, NB, and RF. The deep learning models used were BERT, RoBERTa, and GPT2. The results of the experiment are summarised in Table 2. The performance of the models was evaluated using classification accuracy, the F1-score, and the kappa metric (explained in the previous section). The results show that deep learning models outperformed traditional machine learning models in terms of F1-score and kappa value. GPT2 achieved the best F1-score and kappa values of 0.924 and 0.781, respectively.

The NB model exhibited the lowest performance, achieving an F1-score of 0.355 when utilising Word2Vec features. Similarly, LR demonstrated the least favourable kappa score of 0 when employing the Word2Vec features, suggesting that its classification results were not significantly different from those that could be attributed to random chance alone. In addition, the results indicate that the choice of feature representation also plays a crucial role in the performance of the models. The results demonstrate that the traditional machine learning models showed better performance with TF-IDF features.

Overall, the results of this study suggest that deep learning models like BERT, RoBERTa, and GPT2 are more effective than traditional machine learning models for the task of classifying discourse data into negotiation and non-negotiation. The choice of feature representation also plays a significant role in the performance of the models.

6. Discussion and Conclusion

In conclusion, this study proposes an approach for automating the identification of negotiation in learners' discourse during collaboration. By leveraging language models like BERT, RoBERTa, and GPT, we accurately identified negotiation utterances in learners' discourse while they solved complex engineering estimation problems in an OELE called MEttLE with high precision and recall.

The practical applications of our research have the potential to impact educational settings greatly. We envision the utilisation of language-based collaboration analytics (CA) models in authentic educational environments for targeted interventions to enhance CPS skills. One key application involves the deployment of automated reports to teachers who oversee multiple groups of students engaged in CPS activities. These reports, accessible
through a teacher dashboard, could provide insights into each group's engagement with various aspects of CPS, such as constructing shared knowledge. This functionality empowers teachers to identify groups requiring additional support and allocate their resources effectively. Moreover, teachers can pinpoint individual students' strengths and weaknesses, facilitating the establishment of personalized improvement goals.

Furthermore, the proposed approach isn't limited to teacher-facing analytics; it can also provide learner-facing feedback to foster CPS skill development. CA models could offer individual team members insights into their contributions and demonstration of different CPS skills. This personalized feedback enhances self-awareness, self-reflection, and the evaluation of strengths and weaknesses among learners. Such feedback facilitates tracking skill improvement across multiple collaborative engagements.

Nonetheless, our study is not without limitations. Our results suggest that our approach can accurately identify negotiation utterances in this specific context, but the generalizability of our method to other learning environments and contexts needs to be further investigated. Additionally, the study relied on language models to identify negotiation, which may not capture nonverbal cues and context-specific language use. Future research could explore the use of multimodal methods to identify negotiation in discourse.

Acknowledgements

The authors would like to thank the volunteers for their participation in the study. The authors acknowledge the Research Scholars of the Indian Institute of Technology (IIT) Bombay for their suggestions and input; and the Indian Ministry of Education for financial support in conducting this research.

References


Actions speak loudly with words: unpacking collaboration around the table. In Proceedings of the ACM international conference on interactive tabletops and surfaces (pp. 189-196).


KHWAJA, U., & MURTHY, S. Understanding Learners' Negotiation Processes During Ill-Structured Engineering Estimation Problem-Solving.


A page jump recommendation model based on digital textbook contents and student log data

Wenhao WANG a, Natsumi YAMAMOTO a, Fuzheng ZHAO a, Etsuko KUMAMOTO b, Zicheng KANG c, Chengjiu YIN d

a Graduate School of System Informatics, Kobe University, Japan
b Information Science and Technology Center, Kobe University, Japan
c Faculty of Systems Science and Technology, Akita Prefectural University, Japan
d Research Institute for Information Technology, Kyushu University, Japan

Abstract: In the analysis of student learning data, researchers found that students often switch pages but cannot accurately find the target page. Therefore, in this study, a page recommendation model to provide students with the target page for jumping to improve the efficiency of student page jumping behavior was developed. The recommendation model linearly combines the content-based recommendation model using term frequency-inverse document frequency (TF-IDF) and the recommendation model based on student log data, and determines the optimal weight of linear combination through gradient descent. Finally, the accuracy of the model was evaluated by comparing it with expert-generated recommendation outcomes.

Keywords: Page jump recommendation, TF-IDF, Log data

1. Introduction

Online learning has been on the increase in the last two decades (Martin, 2020). As an important part of online learning, e-books, and other digital teaching materials are gradually replacing traditional paper textbooks due to their low cost and better portability (Wang J, Shimada A, 2023). However, research has found that some students can remember less content when reading textbooks on screens than when reading paper textbooks, and their understanding of the content is shallower than when reading paper textbooks (Lauterman T, Ackerman R, 2014). At the same time, the sudden outbreak of COVID-19 has forced schools around the world to adopt online teaching methods on a large scale. Research by Emma Dorn and Bryan Hancock (2020) shows that COVID-19 has caused significant disruption to the education system and students’ academic performance has declined to varying degrees.

Researching and improving online learning has become an important factor in improving students’ online learning outcomes (Garcia-Morales, 2021). In the analysis of online learning behavior, Yin et al. (2019) found that students with better academic performance have a significantly higher Backing Track Rate (BTR) through the analysis of learning behavior patterns based on student behavior data. Havercamp (2020) also showed through research that backtracking behavior has a uniquely positive effect on improving content understanding when reading digital materials. However, in the analysis of learning data from 122 students, we noted that more than 40% of backtracking behavior requires more than two jumps to navigate to the target page, indicating that students’ backtracking behavior is inefficient.

To improve the efficiency of backtracking behavior and provide students with navigation for page jumping, this study proposes a page recommendation model that combines content-based and user data-based methods to help students find the target page for jumping and improve learning efficiency when reading e-books.
2. Literature review

2.1 TF-IDF algorithm

TF-IDF is a commonly used weighting technique for information retrieval and text mining, which can reflect the importance of a word to a document or a corpus. The TF-IDF algorithm was first proposed by Sparck Jones (1972) to evaluate the performance of information retrieval systems. Later, it is widely used in various natural language processing tasks. For example, Wu et al. (2008) applied TF-IDF to relevance decision, Shahzad Qaiser et al. (2018) used the TF-IDF algorithm to examine the relevance of keywords and documents in the corpus, Zhou et al. (2020) applied TF-IDF to cluster news. Due to its simple mathematical calculation formula, low computational complexity, and relatively good accuracy in classification (Wu, 2018) this study also uses this algorithm in the recommendation model.

2.2 student log data analysis in Education

With the development of web-based education, educational data mining has become a promising field (Romero, 2007), and many studies have been conducted on mining student log data. In mining student log data, Yin et al. (2017) mined learning behavior patterns of different types of students from data of students reading digital books; Zhao et al. (2021) predicted students' academic performance using 6 prediction algorithms based on student data from an e-book system; Riestra-González et al. (2021) predicted student performance in the early stages of a course by analyzing log data from an LMS system; Gobert (2013) evaluated students' abilities in science inquiry through their log data.

In the research direction of recommending something to students based on student log data, Aher (2012) recommended online courses through Moodle system's student course browsing data; Reddy (2016) recommended the learning sequence of courses based on the courses and students' grades.

These studies all focus on the personalized course or learning path recommendations for students, while this paper focuses on recommending pages of digital books at a more specific level.

3. Design and development of page jump recommendation model

In this section, the specific calculation method of TF-IDF in content-based recommendation models, the method of mining page relevance from student log data, and the construction method of hybrid models are mainly introduced.

3.1 Finding page associations based on textbook content

This study uses TF-IDF and cosine similarity to evaluate the relevance of pages. TF-IDF is an analytical method for evaluating the importance of words in a document. TF represents the frequency of a word in a document, and IDF represents the uniqueness of a word in a document (Ramos J, 2003). By combining the two, TF-IDF takes into account both the number of occurrences and the unique value of a word, providing a more comprehensive evaluation of the association between a word and a text. The TF-IDF value is calculated using the following formula:

\[
TF(E|t, p) = \frac{\text{Number of word } t \text{ in page } p \text{ in document } E}{\text{Total words of page } p \text{ in } E}
\]

\[
IDF(E|t) = \log_2 \left( \frac{pc}{\sqrt{\sum(E|t)}} \right)
\]

\[
TF-IDF(E|t, p) = TF(E|t, p) \times IDF(E|t)
\]
Table 1. Formula of TF-IDF

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>E-book number</td>
</tr>
<tr>
<td>pc</td>
<td>Total number of pages of E</td>
</tr>
<tr>
<td>p</td>
<td>A page in E</td>
</tr>
<tr>
<td>t</td>
<td>A word</td>
</tr>
<tr>
<td>(f(E</td>
<td>t))</td>
</tr>
</tbody>
</table>

Through the TF-IDF value of words, each page of the textbook can be simply vectorized, and then the cosine similarity between texts can be calculated. The closer the cosine similarity between two vectors is to 1, the more similar the texts are. Repeating the calculation process we can get a similarity matrix \(R_a\) between texts.

### 3.2 Mining page relevance based on students’ learning data

According to the research results of Chiba (2017), the reading speed evaluation index words per minute (wpm) for students who use English as a second language when reading English materials is 144.42 wpm. Based on this wpm, the time \(T_r\) required for a student to read a page can be inferred from wpm and the total number of words on the page \(W_s\) by this formula:

\[
T_r = \frac{W_s}{wpm/60s}
\]

When a student performs a jump page behavior, a page sequence will be generated. Considering that the student will not read all the content on the target page completely, we use 0.8 times the reading time \(T_r\) of that page as the standard duration for that page.

![Figure 1. The method of judging page relevance.](image)

When the student’s stay time on a certain page in the page sequence is greater than or equal to the standard duration of that page, it is considered that the student has found the target page of this jump, that is, the page is related to the starting page of the jump behavior, and it is reflected in the text similarity matrix \(R_\delta\) as an increase of 1 in the corresponding element value. Then each element is divided by the sum of the elements in its row and normalized to update the matrix \(R_\delta\).
3.3 The construction of the recommendation model

This study combines the content-based text similarity matrix $R_\alpha$ and the user data-based text similarity matrix $R_\beta$ linearly according to the following formula to obtain the text similarity matrix $R(\alpha, \beta)$ of the hybrid model, where $\alpha$ and $\beta$ are obtained by gradient descent. Then sorting each row of the $R(\alpha, \beta)$ matrix in descending order we can obtain the recommendation list for each page.

$$R(\alpha, \beta) = \alpha R_\alpha + \beta R_\beta$$

This paper uses the following formula (Bartell, 1994) to find the optimal values of the parameters $\alpha$ and $\beta$.

$$J(\alpha, \beta) = \frac{\sum_{j > q, k} [R_{ij}(\alpha, \beta) - R_{ik}(\alpha, \beta)]}{\sum_{j > q, k} [R_{ij}(\alpha, \beta) - R_{ik}(\alpha, \beta)]}$$

$$\arg \min J = \frac{1}{|Q|} \sum J_i(\alpha, \beta)$$

$j > q, k$ means that for page $i$, the user is more likely to jump to page $j$ than to page $k$. $R_{ij}$ is the similarity value between page $i$ and page $j$, while $R_{ik}$ is the similarity value between page $i$ and page $k$. $Q=\{q_i\}$ is the recommendation list for the textbook made in advance by experts, where $q_i$ is the recommendation list for page $i$. When the order of the recommendation list generated by the model is completely consistent with the order of the expert recommendation list, $J=1$.

By comparing the order of the recommendation list generated by the model with the expert-generated recommendation list, the $\alpha$ and $\beta$ that minimize $J$ are the optimal parameter values.

4. Model evaluation and result

4.1 Model evaluation method

In 2.3, part of the expert-generated recommendation list is used as the training set and the recommendation list $R_m$ generated by the hybrid model is determined. Another part of the expert-generated recommendation list is used as the validation set $R_e$. When one of the top three recommended pages given by $R_m$ is included in the top three recommended pages of $R_e$, the model recommendation is considered successful. The evaluation index accuracy is obtained by calculating the quotient of the number of successful pages and the total number of pages.

4.2 Model evaluation result

In this study, we collected data from 122 university students who use English as a second language studied the “Commercial Law” digital textbook in English and mined the associations between texts from this data.

![Figure 3. The text image of “Commercial Law”.](image-url)
Table 2. Samples of log data

<table>
<thead>
<tr>
<th>User ID</th>
<th>Operation</th>
<th>Date</th>
<th>Page No</th>
</tr>
</thead>
<tbody>
<tr>
<td>admin</td>
<td>NEXT id:12</td>
<td>2016/10/12</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>page:39</td>
<td>14:00:31</td>
<td></td>
</tr>
<tr>
<td>demo</td>
<td>NEXT id:12</td>
<td>2016/10/12</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>page:45</td>
<td>14:01:05</td>
<td></td>
</tr>
<tr>
<td>admin</td>
<td>PREV id:12</td>
<td>2016/10/12</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>page:41</td>
<td>14:00:34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bookpage:41</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Through gradient descent, the final determined values of $\alpha$ and $\beta$ are 135.54422448513 and 116.968136487428 respectively. The hybrid model, the recommendation model based on user data and the content-based recommendation model were evaluated according to the method in 3.1, and the results are shown in the table.

Table 3. The accuracy of models

<table>
<thead>
<tr>
<th>Category</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hybrid model in this study</td>
<td>0.64</td>
</tr>
<tr>
<td>user data-based model only</td>
<td>0.61</td>
</tr>
<tr>
<td>content-based model only</td>
<td>0.37</td>
</tr>
</tbody>
</table>

The accuracy of the recommendation list generated using only user data is 0.61, the accuracy of the recommendation list generated using only textbook content is 0.37, and the accuracy of the model linearly combined through gradient descent reaches 0.64. The results prove that the recommendation list generated by the hybrid model is more accurate and effective than using only teaching material content or user data analysis alone.

5. Conclusions

This study linearly combined the recommendation model based on textbook content and the recommendation model based on user data, and determined the optimal weights of the two through gradient descent, resulting in a hybrid recommendation model for the page. After comparing with expert data, the accuracy of the hybrid model was evaluated, proving the accuracy of the page recommendation model is higher than that of a single model.

On the other hand, the poor performance of the recommendation model based solely on content indicates that there is still much room for improvement in the processing methods of the textbook content. And due to the lack of expert-generated recommendation lists and student learning logs, the overall performance of the model is not satisfactory.

Therefore, more research is needed to improve the recommendation model, and the effectiveness of the recommendation system still needs to be verified in actual learning environments.

Acknowledgements

This research was partially supported by Grants-in-Aid for Scientific Research No. 21H00905 from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan.

References


Concept Map Recomposition Approach for Advanced Formative Assessment in Large-Scale Online Course

Sho YAMAMOTO\textsuperscript{a}, Aryo PINANDITO\textsuperscript{b} & Tsukasa HIRASHIMA\textsuperscript{c}

\textsuperscript{a}Faculty of Informatics/ Cyber Informatics Research Institute, Kindai University, Japan

\textsuperscript{b}Faculty of Computer Science, Universitas Brawijaya, Indonesia

\textsuperscript{c}Graduate School of Advanced Science and Engineering, Hiroshima University, Japan

*yamamoto@info.kindai.ac.jp

Abstract: Assessment of learner understanding and feedback based on that assessment are important to promote learning in asynchronous online learning. Often, automatic feedback by the learning environment is used as a way of realizing this feedback. However, usual tests are difficult to identify learning objects exhaustively. Concept maps can provide a comprehensive and detailed check of a learner’s understanding but are difficult to assess. Therefore, this study used a recomposition concept map called Kit-Build concept map (KB map). KB map system can provide semantic feedback to the learner on differences in understanding automatically. We used this feedback as a formative assessment for the entire lecture. Using the KB map as an assignment for a formative assessment, we conducted a large-scale asynchronous course with a class of 1,380 students from 16 faculties. The results show that learners’ understanding is enhanced by concept maps, and that formative assessment by means of concept maps may encourage revision of lesson materials.

Keywords: Formative assessment, asynchronous online learning, assignment, kit-build concept map

1. Introduction

Since the COVID-19 outbreak, online learning has spread rapidly and divergently, and the demand for effective online learning has become an increasingly important issue. There are two types of online learning methods, namely synchronous and asynchronous. In asynchronous online learning, the teacher prepares teaching materials and provides them to learners, who then study at their own pace and time. The asynchronous type is the subject of this study. This type of online learning allows learners to manage their own learning and provides personalized learning opportunities (Garrison, 2003; Nguyen, 2015).

On the other hand, some learners do not have sufficient opportunities to interact during learning (Thamri, Hasan, et al, 2022). Formative assessment is known to help these learners to adjust their own learning (Meusen-Beekman, Joosten-ten, and Boshuizen, 2016). Formative assessment is broadly defined as feedback provided during the learning process (Dunn & Mulvenon, 2009). However, it is also known that this is difficult to realize (Gulikers, Veugen, and Baartman, 2021).

In facilitating appropriate learning for learners, we felt that it was important to be able to assess comprehensively and semantically against the content being taught, and KB maps uses concept maps to externalise learners’ understanding and provides automatic semantic feedback on the difference in understanding with the teacher (Hirashima, Yamasaki, et al, 2015). The possibility of such personalized auto-assessment means that personalized formative assessment is also possible in asynchronous online learning, which may promote learners’ reflection on the teaching material. This study examines the effectiveness of formative assessment using KB maps in asynchronous online lectures. The lecture is an
2. Related Works

Formative assessment is defined as feedback provided during the learning process (Bennett, 2011). Formative assessment is useful for learners to self-regulate their learning, it is difficult to achieve appropriate formative assessment (Gulikers, Veugen, and Baartman, 2021). For appropriate formative assessment, it is important to provide feedback on understanding of the learning subject in a form that is both comprehensive and easy for learners to understand in order to progress their learning. Various automated systems have been proposed to achieve this appropriately and at a low cost. They are often primarily automated for scoring general tests (Febriani & Abdullah, 2018). However, even when test results are given as scores, it is difficult for learner to reflect on s/he's understanding. Tests also make it difficult to comprehensively assess what learners have learnt. Contrastingly, we focused on formative assessments and self-regulated learning using concept maps. A concept map externalizes the structure of a target concept by connecting concepts as nodes and their relationships as links (Novak & Cañas, 2006). Concept maps can provide a comprehensive and detailed check of a learner's understanding and provide semantic feedback. Thus, it is useful to understand the learner's state of understanding more directly and has been suggested as an effective tool for formative assessment (Beaudry & Wilson, 2010).

However, concept maps are difficult to assess automatically, and they incur high costs for map assessment. They often operate such that teachers and peers view the learner's map and interact with it for learning. It is difficult to carry out such activities in asynchronous online learning. We therefore turned to KB maps. In contrast, KB maps are used to automatically assess learner maps (Hirashima, Yamasaki, et al., 2015). KB maps allow the comparison of maps constructed by the teacher and the learner to provide automatic feedback on the differences between each map as errors or lack of understanding of concepts by the learner.

The validity of the automatic assessment of KB in comparison with several manual methods has already been confirmed (Wunnsari, Pailai et al., 2018). The Kit-Build concept map system (KB Map System) has been used in face-to-face learning for the formative assessment of each lesson and has been shown to enable teachers to provide more relevant feedback to learners (Pailai, Wunnsari, et al., 2017). Therefore, we considered that integrating KB Maps into asynchronous online classes would contribute to building a cycle in which learners could deepen their own learning.

3. Kit-Build Concept Map System for Asynchronous Online Classes

A procedure of the formative and summative evaluations based on the KB map used in this study is shown in Figure 1. The teacher represents summarized concepts that s/he wants to transfer to the learner as a concept map (summarized map). The summarized map can be created on a KB map system and automatically decomposed into a set of links and nodes by a KB map system. The decomposed map is referred to as a kit, were created for each class.

For each class, the learner downloads a kit created by the teacher and build it to represent his/her understanding of the taught content (learner map) as assignment. Feedback is available at any time for each assignment. The learner can press the feedback button during map building to determine whether there are links that are improperly connected. In this study, assessment carried out at this time is called formative assessment. In this case, incorrect or unconnected concepts are highlighted using untagged links.

In asynchronous online classes, learners are free to attend classes at any time, making it difficult for teachers to directly intervene in learners' learning. However, the KB Map system provides real-time and semantic feedback based on concepts and its relationships, so that learners can reconsider the lesson content accordingly, leading to a higher quality their understanding of each lecture. After a series of individual lessons, the learners completed a
KB map-building test and a fill-in-the-blank test to check their understanding of the lesson. In this study, the assessment of these tests is referred to as summative assessment. The results of formative assessment are provided to learners as feedback, whereas those of summative assessment are only used by teachers.

Based on the above, the following are the research questions (RQ) for this study as an initial evaluation of the formative evaluation framework using the KB Map: (RQ1) Does formative assessment using the KB maps improve summative assessment performance? (RQ2) Does formative assessment using KB maps encourage positive learner behavior?

![Figure 1. Framework for formative and summative assessment using KB map.](image)

4. Design of the Online Course

4.1 Procedure of Practical Use

This practical application was targeted at a lecture called “Introduction to Data Literacy.” The classes discussed in this paper are lectures 1st, 3rd to 9th, which were delivered in the format described in chapter 3. The participants were 1380 first-grade university students from 16 different faculties. Data from 1113 participants were analyzed in this study. There were three reasons for adopting the log data: (1) Participants agreed to use the data, (2) The feedback from the KB Map system is not an outlier, and (3) complete data. The data were handled in compliance with the ethical codes established by the first author’s university.

Contents of 1st, 3rd to 8th lectures will explain. The target classes are asynchronous online lectures with weekly instructional videos and students may view the videos at their own pace and timing. However, learners were required to view videos within two weeks. After viewing the instructional videos, students work on KB Map assignments based on the instructional videos, which are accepted until two weeks after the end of the 15th class. However, students are informed that points will be reduced by 50% after two weeks—the deadline for the videos.

The second lecture was in an omnibus class in which the learner selected several of the video materials to take; thus, the KB Map assignment was not conducted. In the 9th class, a KB map test and a fill-in-the-blank test were conducted; the KB map test consists of the concepts used in the KB map assignments from classes 1, 3 to 8. The fill-in-the-blank test is provided in the form of a text using these concept. The fill-in-the-blank test is a multiple choice question; all answers are presented as choices for each question. In addition, the KB Map and fill-in-the-blank tests each had a 30-minute time limit.

4.2 Scale of Concept Maps and Assignments for Each Class

A concept map was created by extracting and connecting concepts from slides that served as class materials. The KB Map does not strictly follow the notation defined in the concept map, but the emphasis is on students and teachers agreeing with the concepts being represented. The scale of the KB Maps that the learners were asked to build was such that the assignments had an average number of 39.57 nodes and 39.14 links. Note that some of the nodes and
links were pre-built, so the actual number of nodes averaged 33.42, and the number of links was 32.71. In KB map test, the nodes and links were 42 and 41, respectively. Excluding the pre-built map, there were 23 nodes and 22 links. Two class co-tutors approved the concepts used in the map.

Learners can use immediate feedback while building maps to deepen their understanding of poorly understood concepts and relationships. However, because this feedback is not presented until the links that should be used to connect them are determined, the learner is relearned to confirm his/her understanding of the necessary concepts and their relationships by, for example, referring to the class video.

5. Results and Discussion

5.1 Data Acquisition

The data used in this analysis are the log data of the KB Map, the scores of the fill-in-the-blank test, and the log data of viewing online videos. The log data of the KB Map are the map scores and feedback counts for the KB map assignments. The log data for the KB Map test is only the map score. The fill-in-the-blanks test scores were also used by converting the scores of all 21 questions into percentages of 100. From the viewing logs of online video materials, the number of times the speed of the video is changed, the number of times the viewing point of the video is changed (seek bar operation) and the number of times the video is stopped can be obtained. For our analysis, we divided these data by the number of classes and used them as the average number of times per class. The map score, fill-in-the-blanks test score, and assignment score were scored on a 100-point scale, but other parameters were not capped.

5.2 Results

The means and standard deviations for each data point are listed in Table 1. The average assignment score was 94, with a submission rate of 95%. The average number of times the feedback was actively used was 90. The fact that many students were active despite the KB Map was a new assignment suggests that it was accepted as an assignment.

Multiple regression analysis (MRA) was used to determine the relationship between each test or log. Note that for all results in Table 2 to 5, the Signif. codes are denoted as: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 1. From the result of Table 2, there appeared to be a relationship between scores on the fill-in-the-blanks test and the KB map test, as well as scores on the tasks. However, the multiple R-squared value was 0.06, which is not highly explanatory. Spearman’s correlation analysis was conducted between the KB Map and fill-in-the-blank test scores. The results showed a slightly positive correlation (r = .23). There was also little correlation between fill-in-the-blank test scores and the number of feedbacks (r=.11).

The result of Table 3 showed that both the number of feedbacks and the average score on the KB Map assignment have an effect. However, the multiple R-squared value was low (.17). Therefore, Spearman’s correlation analysis was conducted on the map test scores and feedback counts; however, there was no correlation (r = -.02). The average score on the KB Map assignment was positively correlated with the average score on the KB Map test (r = .40).

<table>
<thead>
<tr>
<th>Table 1. Mean and Standard Deviation of Each Collected Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
</tr>
<tr>
<td>Average score of KB map test</td>
</tr>
<tr>
<td>Average score of fill-in-the-blank test</td>
</tr>
<tr>
<td>Average score of KB map assignment</td>
</tr>
<tr>
<td>Average number of feedbacks</td>
</tr>
<tr>
<td>Average number of seek bar operations</td>
</tr>
<tr>
<td>Average number of changing video speed</td>
</tr>
<tr>
<td>Average number of stop operations</td>
</tr>
</tbody>
</table>
The results of Table 4 indicated that the number of feedback sessions had an effect. However, the multiple R-squared value was low (.06). Spearman’s correlation coefficient was calculated, and no correlation was found ($r = .17$). The results of Table 5 showed that feedback counts were related to the number of seek bar operations, speed changes, and stop operations. The multiple R-squared value was low (.02). When the correlation coefficients were determined, it was found that the feedback count and seek bar operations were quite positively correlated, with $r = .41$. The number of stopped operations was also positively correlated ($r = .46$). However, there was little correlation with the speed change ($r = .13$).

### Table 2. MRA results with the objective variable as the fill-in-the-blank test score and the explanatory variables as the KB Map test score, number of feedbacks, and KB Map assignment score.

|                     | Estimated Std. | Error  | t value | Pr(>|t|) |
|---------------------|----------------|--------|---------|----------|
| (Intercept)         | 45.63          | 3.34   | 13.67   | <2e-16 *** |
| Score of KB Map test| 0.22           | 0.04   | 5.76    | 1.11e-08 *** |
| Number of feedbacks | 0.02           | 0.01   | 3.41    | 0.0007 *** |
| Score of KB Map assignment | 0.06 | 0.04 | 1.44 | 0.15 |

### Table 3. MRA results with the objective variable as the KB Map test score and the explanatory variables as the number of feedbacks, and KB Map assignment score.

|                     | Estimated Std. | Error  | t value | Pr(>|t|) |
|---------------------|----------------|--------|---------|----------|
| (Intercept)         | 35.12          | 2.41   | 14.56   | <2e-16 *** |
| Number of feedbacks | -0.03          | 0.005  | -6.76   | 2.2e-11 *** |
| Score of KB Map assignment | 0.58 | 0.026 | 22.39 | <2e-16 *** |

### Table 4. MRA results with the objective variable as the KB Map assignment score and the explanatory variables as the number of feedbacks, seek bar operations, speed change operations, and stop operations.

|                     | Estimated Std. | Error  | t value | Pr(>|t|) |
|---------------------|----------------|--------|---------|----------|
| (Intercept)         | 90.59          | 0.74   | 121.77  | <2e-16 *** |
| Number of feedbacks | -0.001         | 0.01   | -0.09   | 0.93     |
| Number of speed change operations | -0.11 | 0.20 | -0.55 | 0.59 |
| Number of stop operations | -0.03 | 0.05 | -0.52 | 0.60 |

### Table 5. MRA results with the objective variable as the number of feedback and the explanatory variables as the number of seek bar, speed change, and stop operations.

|                     | Estimated Std. | Error  | t value | Pr(>|t|) |
|---------------------|----------------|--------|---------|----------|
| (Intercept)         | 63.98          | 3.31   | 19.35   | <2e-16 *** |
| Number of seek bar operations | 0.29 | 0.06 | 5.08 | 4.39e-07 *** |
| Number of speed change operations | 1.67 | 1.04 | 1.60 | 0.11 |
| Number of stop operations | 2.12 | 0.27 | 7.74 | 2.19e-14 *** |

5.3 Discussion and Limitation

Analysis results indicate that a deeper understanding of the KB Map test may also improve scores on the fill-in-the-blank test. The KB Map test scores are also related to the KB Map assignment scores, with a deeper understanding of KB Map assignments also improving the KB Map test scores. This is a reasonable result because the concept of KB Map assignment was used in the KB Map test. However, while we found that the KB Map assignment scores were related to the number of feedbacks, we did not find any correlation. One possible reason is that more frequent feedback does not necessarily lead to more effective learning. Therefore, although the possibility of meeting RQ1 was identified, further analyses are necessary.
In addition, analysis results suggest that feedback may have encouraged learners to engage in activities to review the class video again. This confirmed the possibility that formative assessment of the KB Map assignment in the asynchronous online class using feedback could point out the learners' lack of understanding and allow them to recheck the material, thereby enabling them to acquire concepts that the teacher wanted to transfer through appropriate trial and error. These results indicate that RQ2 was identified. However, a detailed analysis of the feature of learners and how feedback is used remains a challenge.

6. Conclusion and Future Works

In this study, we propose the use of a KB Map and framework to realize effective formative assessment in asynchronous online classes. This formative assessment should be comprehensive and able to point out sources of error against the learning content in order for the learner to reflect on his or her own learning and to be effective. Therefore, in this study, we evaluate asynchronous online learning using the KB Map System, which automatically evaluates learners’ concept maps.

Consequently, it was confirmed that formative evaluation in the KB Map system may reaffirm the learning material. We were also able to confirm that the performance of the KB Map building influenced the final summative evaluation, and we believe that we were able to provide a useful cycle of asynchronous online learning to learners. However, we have not yet been able to examine how formative assessment was utilized and how the effects differed depending on the attributes of the learners; therefore, a detailed evaluation is needed.

References


Towards Automated Evidence Extraction: A Case Study of Adapting SAM to Real-World Educational Data

Kouki Okumura\textsuperscript{a}, Izumi Horikoshi\textsuperscript{b}, Kento Koike\textsuperscript{b} & Hiroaki Ogata\textsuperscript{b}
\textsuperscript{a}Graduate School of Informatics, Kyoto University, Japan
\textsuperscript{b}Academic Center for Computing and Media Studies, Kyoto University, Japan
*okumura.kouki.27m@st.kyoto-u.ac.jp

Abstract: The demand for a shift from intuition- and experience-based education has been growing. A major challenge in realizing this transition is extracting evidence from real-world educational data. Conventionally, this data extraction process is performed by manually choosing classes for comparison. However, the selection requires expert knowledge of which classes should be compared with which indicators. In this study, we propose the use of deep learning algorithms to uncover inherent causal relationships within vast amounts of existing data. Specifically, we employ structural agnostic modeling, a causal search algorithm known for its exceptional performance on real-world data, to extract complex causal candidates in education. This approach has been referred to as “observational causal discovery.” We evaluate the effectiveness of this method using real-world educational data and compare its advantages with those of conventional automatic comparison methods. Results demonstrate that the proposed method can identify various causal candidates, including those that are difficult for humans to discern, and even those without causal relationships. Different from existing methods, the proposed approach does not require selected or fixed comparative indices, thus potentially uncovering comparative indices that elude human comprehension. We anticipate that this research will enable the collection of substantial evidence from real-world educational data and promote evidence-based education.

Keywords: Evidence-based education, real-world evidence, SAM, causal search, causal analysis

1. Introduction

Evidence-based education (Davies, 1999) is expected to provide education that is not based on intuition or experience. Evidence is generally extracted from systematic reviews and randomized controlled trials (RCTs). Although these methods provide high-level evidence, collecting large amounts of evidence is difficult because doing so is often costly, unethical, or unfeasible (Slade & Prinsloo, 2013). Therefore, the concept of real-world evidence has been proposed in the medical field (Mahajan, 2015), and several attempts have been made to apply it to the educational field as well. Real-world evidence is evidence extracted from various real-world data, and although it is less reliable than that extracted from systematic reviews and RCTs, it is expected to be voluminous.

Our previous method compares a control group with a corresponding target group to verify the effects of an intervention (Nakanishi, 2021). This method is advantageous as it allows the effects of an intervention to be verified by setting target and control groups, thereby eliminating the need for experiments. However, it requires manually setting the classes to be compared, and its use in automatic extraction from many cases is difficult.

To address this issue, we developed a method that can automatically search for classes with similar contextual information, such as grade level, as a control group (Okumura et al., 2022). However, this method allows only one factor to be considered and compared. To further update this method, we adopted a method called “observational causal discovery.”
Observational causal discovery is the process of hypothesizing causal relationships from observational data. This method has attracted attention from the machine learning community, thus reflecting the difficulty of conducting RCTs, which are the gold standard for establishing causal relationships (Lopez-Paz et al., 2015; Mooij et al., 2016; Peters et al., 2017). Observational causal discovery has been applied to many domains, such as in economics to understand and model the impact of monetary policy and in bioinformatics to infer network structures from gene expression data and prioritize exploratory experiments (Kalainathan et al., 2018, 2022). However, we have yet to find many applications of observational causal discovery in educational data. Therefore, we raise the following two research questions (RQs):

- **RQ1**: What do we obtain from the observational causal discovery method adapted to real-world educational data?
- **RQ2**: How different are the results of the conventional method and observational causal discovery?

To address the above RQs, we attempt to apply the observational causal discovery method, which has already been applied in other fields, to real-world educational data.

For RQ1, we apply the causal analysis algorithm called structural agnostic modeling (SAM), which has shown great performance in real-world biological data, to real-world educational data and verify its effectiveness. For RQ2, we compare our previously proposed method with the method using SAM and clarify their characteristics and differences. Through these RQs, we aim to present a new path for more effective evidence extraction from educational data.

## 2. Method

### 2.1 Overview

The methodology used in this study consists of two steps. The datasets used are listed in Table 1. The first step is for RQ1, and it involves examining whether SAM is applicable to educational data and whether it provides effective results. Specifically, we select dataset **Y** for analysis and apply SAM to it. The second step is for RQ2, and it involves comparing the SAM-based method with our previously proposed method. For this step, we apply SAM to dataset **Z**, which is from dataset **X** for the same activity as that in the previous study by Okumura et al. (2022).

In Japan, one tablet or a similar device is distributed to each student under the GIGA School Project. This project uses a learning analytics platform called the LEAF system (Ogata, 2018) and collects the log data accordingly. In the current study, we utilize the log data from the LEAF system.

The analysis is conducted as follows (Figure 1): First, the log data containing information on the students’ activities in the platform, including what the activities are and when they are completed, is collected. Second, this information is summarized to show the number of operations with the digital textbooks in each class. Third, a causal search is conducted to determine whether a causal relationship exists. Fourth, the results are displayed as a table in the form of 0s and 1s according to a set threshold. Finally, the results are output as a directed acyclic graph (DAG).
Table 1. Comparison of Our Approach and Conventional Studies.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Analysis Method</th>
<th>Data source</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Manual Comparison</td>
<td>a junior high school</td>
<td>Nakanishi (2021)</td>
</tr>
<tr>
<td>Dataset X</td>
<td>Automated Comparison</td>
<td>a junior high school</td>
<td>Okumura et al. (2022)</td>
</tr>
<tr>
<td>Dataset Y</td>
<td>Automated Causal Exploration (SAM)</td>
<td>a university</td>
<td>This study</td>
</tr>
<tr>
<td>Dataset Z</td>
<td>Automated Causal Exploration (SAM)</td>
<td>a junior high school</td>
<td>This study</td>
</tr>
</tbody>
</table>

2.1.1 Dataset from a University (Dataset Y)

We used the data collected from an educational data analysis contest conducted by the Council for Evidence-Driven Education Research (EDE, 2022).

Specifically, we utilized the educational data obtained with the LEAF system at a university. These data included digital educational material browsing behavior in four courses offered as Kyu data in the 2019 and 2020 academic years (Table 2). The logs were organized according to class time and content IDs, and data were produced for each indicator for each class. The indicators are listed in Table 3.

Table 2. Summary of Dataset Y.

<table>
<thead>
<tr>
<th>Course Id</th>
<th>Term</th>
<th>Student</th>
<th>Log Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-2019</td>
<td>2019 Spring</td>
<td>50</td>
<td>129,358</td>
</tr>
<tr>
<td>B-2019</td>
<td>2019 Spring</td>
<td>164</td>
<td>352,467</td>
</tr>
<tr>
<td>A-2020</td>
<td>2020 Spring</td>
<td>62</td>
<td>147,452</td>
</tr>
<tr>
<td>B-2020</td>
<td>2020 Spring</td>
<td>93</td>
<td>197,593</td>
</tr>
</tbody>
</table>

Figure 1. Method Overview.
Table 3. Indicators from Dataset Y.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNT LP</td>
<td>Count of Lecture Period</td>
</tr>
<tr>
<td>SUM OP</td>
<td>Sum of Operations</td>
</tr>
<tr>
<td>AVG OP</td>
<td>Average Operations</td>
</tr>
<tr>
<td>SUM ATT</td>
<td>Sum of Attending Students</td>
</tr>
<tr>
<td>PG LP</td>
<td>Number of Pages Covered in Lecture Period</td>
</tr>
<tr>
<td>DV LP</td>
<td>Number of Devices Used in Lecture Period</td>
</tr>
<tr>
<td>SUM MRK</td>
<td>Sum of Markers</td>
</tr>
<tr>
<td>AVG MRK</td>
<td>Average Markers</td>
</tr>
<tr>
<td>NUM MAT</td>
<td>Number of Teaching Materials Used in Lecture Period</td>
</tr>
<tr>
<td>AFE ATT</td>
<td>Average Final Exam Score of Attending Students</td>
</tr>
<tr>
<td>ONLINE</td>
<td>Whether the Course was Online or Not</td>
</tr>
</tbody>
</table>

2.1.2 Dataset from a Junior High School (Dataset Z)

For Dataset Z, we analyzed the logs of the days when active reading activities were conducted at the same junior high school in school years 2021 and 2022. Active reading is “considered best suited for textbook reading, helping learners understand unfamiliar information in texts, and providing a structured approach to learning” (Toyokawa et al., 2023, p. 2). Active reading enables one to read quickly and understand important points, and it involves the use of multiple learning tools, such as e-books and learning analytics tools. The target class consisted of three sessions, namely, Day 1, Day 2, and Day 3. E-book memos, markers, and timers were used in the activities. In addition, quizzes were administered, and reading speed (words per minute, WPM) was measured before and after the class (pre- and post-class). The indicators computed from the log data of these activities are listed in Table 4.

Table 4. Indicators of Dataset Z

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARD D1</td>
<td>Active reading dashboard use on Day 1</td>
</tr>
<tr>
<td>ARD D2</td>
<td>Active reading dashboard use on Day 2</td>
</tr>
<tr>
<td>ARD D3</td>
<td>Active reading dashboard use on Day 3</td>
</tr>
<tr>
<td>MRK D1</td>
<td>Marker use on Day 1</td>
</tr>
<tr>
<td>MRK D2</td>
<td>Marker use on Day 2</td>
</tr>
<tr>
<td>MEM D1</td>
<td>Memo use on Day 1</td>
</tr>
<tr>
<td>MEM NC</td>
<td>Memo use on a day with no class</td>
</tr>
<tr>
<td>MEM D2</td>
<td>Memo use on Day 2</td>
</tr>
<tr>
<td>MEM D3</td>
<td>Memo use on Day 3</td>
</tr>
<tr>
<td>RDG D1</td>
<td>Reading operations on Day 1</td>
</tr>
<tr>
<td>RDG D2</td>
<td>Reading operations on Day 2</td>
</tr>
<tr>
<td>RDG D3</td>
<td>Reading operations on Day 3</td>
</tr>
<tr>
<td>TIM D1</td>
<td>Timer use on Day 1</td>
</tr>
<tr>
<td>TIM D2</td>
<td>Timer use on Day 2</td>
</tr>
<tr>
<td>TIM D3</td>
<td>Timer use on Day 3</td>
</tr>
<tr>
<td>QUZ PRE</td>
<td>Score in the pre-quiz</td>
</tr>
<tr>
<td>QUZ PST</td>
<td>Score in the post-quiz</td>
</tr>
<tr>
<td>WPM PRE</td>
<td>Words per minute before the active reading activity</td>
</tr>
<tr>
<td>WPM PST</td>
<td>Words per minute after the active reading activity</td>
</tr>
<tr>
<td>SUM SMR</td>
<td>Score of summary after active reading class</td>
</tr>
</tbody>
</table>
2.2 Applied Algorithm: SAM

SAM (Kalainathan et al., 2018, 2022) utilizes a type of deep learning called generative adversarial network (GAN). GAN is known for its ability to generate fake objects that are close to real objects by having two networks, namely, a generative network and a discriminative network, compete. SAM is an application of the GAN concept to causal analysis and is used to determine causal relationships among real-world educational data (Figure 2). Specifically, SAM quantifies the likelihood of causal relationships between indicators and examines whether causal relationships exist. The hyperparameters are the same as those used in the original study.

![Figure 2. Algorithm of Structural Agnostic Modeling (SAM) (Adopted from Kalainathan et al. (2018, 2022))](image)

3. Results

3.1 RQ1: What do we obtain from the observational causal discovery method adapted to real-world educational data? (Dataset Y)

To answer RQ1, we examined whether SAM is applicable to educational data and whether it provides effective results. SAM was applied to dataset Y for analysis. Figure 3 shows the result of SAM for dataset Y.

![Figure 3. Result of SAM for Dataset Y.](image)
(1) Intuitively understandable result
The analysis revealed a causal relationship between the number of teaching materials used during class (MV LP) and the average number of operations (AVG OP) (Figure 3, arrow 7). This result is intuitively understandable result: the more types of teaching materials available, the greater the number of operations.

(2) Counterintuitive results
However, a causal relationship was found between the number of device types used in class (DV LP) and the semester grade point average of the student attendees (AFE ATT) (Figure 3, arrow 2). This result counters the intuition as more types of devices lead to higher grades. If this relationship is actually a causal one, then it is a new educational technique worth trying. The causal relationship between whether the course was offered online (ONLINE) and end-of-semester grades (AFE ATT) (Figure 3, arrow 3) was also counterintuitive. The notion of academic grades increasing with the availability of online classes is difficult to envision. However, if such relation is true, then new educational possibilities will become apparent.

(3) Reversed direction causal relationship
A causal relationship was found between the total number of operations (SUM OP) and the number of participating students (SUM ATT) (Figure 3, arrow 4). However, this causal relationship was reversed. The number of operations, of course, did not increase the number of students. Instead, the greater the number of students, the more operations involved; this case shows a more natural causal relationship.

3.2 RQ2: How different are the results of the conventional method and observational causal discovery? (Dataset Z)

To answer RQ2, we applied SAM to dataset Z, which was from the same activity as that in the previous study (dataset X). Figure 4 shows the result of SAM for dataset Z.

![Figure 4. Result of SAM for Dataset Z.](image)
(1) Effects of interventions
The analysis revealed the effects of the interventions we aimed to validate. For example, we found that active reading activities on Day 2 (SUM ATT) affected the quality of memo summaries (SUM SMR) (Figure 4, arrow 3c).

(2) Reversed direction causal relationship
In addition, we noted some reversals of the time axis, which suggested the possibility of causality. For example, we observed a relationship between WPM in the post-class activity (WPM PST) and memo use on Day 1 (MEM D1) (Figure 4, arrow 1). Given the order of the activities, the direction of the causal relationship must be reversed. However, if the direction was adjusted, the causal relationship between memo use and WPM cannot be easily found through human observation. Hence, AI that can process large amounts of data is at an advantage as it is able to suggest the existence of relationships that are not considered relevant by humans.

(3) Absence of causal relationship
The results also suggest not only the “possibility that a causal relationship exists” but also the “possibility that no causal relationship exists.” For example, we did not observe an edge between active reading dashboard use on Days 1 and 3 (ARD D1 and ARD D3). This result indicated that the activities with the dashboard on Days 1 and 3 were not highly effective, whereas those on Day 2 might have been very effective. In future designs for active reading classes, an emphasis on the activities conducted on Day 2 could be considered. If effective, such designs will contribute to improving the learning effectiveness of learners while reducing the burden on teachers.

4. Discussion

4.1 Results of RQs

(RQ1) What do we obtain from the observational causal discovery method adapted to real-world educational data? (Dataset Y)

From these results, we confirmed the possibility of applying SAM to educational data and generating useful results. However, not all causal relationships shown in the DAG were true; therefore, care must be taken when interpreting them. An advantage of this method is that it can process large amounts of data and suggest causal relationships that humans may not be able to assume. Our results also implied the possibility of selecting the causal relationships that should be examined more deeply. In addition, the results indicated that SAM can be useful for educational data analysis. However, maximizing this method requires the proper interpretation and verification of the causal relationships in the output.

(RQ2) How different are the results of the conventional method and observational causal discovery? (Dataset Z)

We compared the automatic causal search method SAM with the previous automatic comparison method and found some differences. Specifically, SAM does not require the indicators to be fixed, whereas the previous method requires the indicators before comparison to be fixed. In this way, SAM may expand the exploration of factor candidates and lead to the possibility of finding comparative indices that humans would not have been able to consider. This feature is an advantage of the automatic causal search method and indicates its potential to open the door for new research.
4.2 Limitations

Addressing RQ2 involved extracting and analyzing the data with background information provided by a researcher who conducted an active reading class study using the LEAF platform. In fully automating the mining process without the need for any manually inputted information, the background information from which the data were obtained and that is tied to the data in some way must be stored to automate the extraction of evidence. We must note that DAG assumes that education is acyclic, although various and complex causal relationships can be hypothesized. SAM produces candidate causal relationships that have the time axis reversed and thus needs improvement so that it does not produce such relationships.

4.3 Implications

Unlike the conventional method, the method in this study was not limited to a single indicator for comparison. Hence, the method proposed herein is able to find comparative indices that humans cannot assume, such as the relationship between memo use and WPM in Section 3.2 (2).

As shown in Section 3.2 (3), knowing specifically which day’s activities are effective may reduce the burden on teachers and improve learning effectiveness.

5. Conclusion and Future Works

This study attempted to apply SAM to the automated extraction of evidence from real-world educational data and examined its effectiveness. The results showed that the causal relationships output by SAM do not necessarily indicate true causal relationships. However, such results are not necessarily meaningless, and the ability to automatically process large amounts of data presents the possibility of various causal relationships that are difficult to identify based on human assumptions.

In addition, through the proposed method, we showed the possibility of proceeding with the analysis even without fixing the indicators for comparison. This feature enables the identification of comparative indices that humans cannot assume. It is also an advantage of using AI and computers as it can provide new perspectives that go beyond those derived by humans.

Future studies should conduct causal inferences in addition to causal search and evaluate the magnitude of causal effects. In addition, they may compare our method with other algorithms for causal search. One of the goals of the current research was to implement the proposed method as a system for teachers and provide real-world evidence of learning. In achieving this goal, the method must be improved and developed while obtaining feedback from teachers based on the assumption that the tool will be used in an actual educational environment.

Acknowledgements

This study was supported by NEDO JPNP20006 and JSPS KAKENHI JP23H00505.

References


What does process mining of feedback-behavior reveal about problem-solving in chemistry undergraduates?

Anveshna SRIVASTAVA*, & Chandan DASGUPTA*

*IDP in Educational Technology, Indian Institute of Technology Bombay, India
*anveshna.sriv@gmail.com

Abstract: Feedback is known to play a powerful role in the process of learning and problem-solving. In a problem-solving context where feedback is expected to positively influence the process of achieving the target goal, feedback information can be elicited through feedback-seeking questions. In this study, we analyze different purposes for which feedback is both sought and given by chemistry undergraduate students (n=3) as they solved representational problems concerning a medicinal drug. The students were left unsupervised and, with their consent, were video-recorded to capture their problem-solving process. We used the transcript of the video data and coded students' interactions to identify feedback events concerning 'the task', 'process', and 'regulation' of the problem-solving activity. These events were then used to generate process models of feedback behavior of students, individually and collaboratively, through ProM- a process mining tool. Our models suggest that feedback questions on tasks and processes dominate the beginning of the problem-solving activity, while feedback responses on processes and regulation dominate the end phase. Our work has implications for designing activity problems. We suggest designing well-thought-out focus questions at the beginning of problem-solving exercises to both scaffold and facilitate students' learning because it aligns with the natural unsupervised flow of problem-solving, as is observed in this study.

Keywords: Feedback, feedback-seeking, questions, responses, process mining

1. Introduction

We seek feedback because, as humans, we are inherently curious and want to know how we are doing and how we have done (Cutumisu, & Lou, 2020). The process of seeking feedback and the corresponding gathered response positively influences learning (Rattan et al., 2015). It helps learners to verify their position, gather insights about strategies that can be used, and/or provide appropriate metacognitive support (Wang & Wu, 2008). In their feedback model, Hattie & Timperley (2007) identified that feedback questions were concerned with tasks, processes, self-regulation, and self, while feedback responses were concerned with answering questions about goal-directed action, overall progression (say, about a concept), and future possibilities. In this study, we draw upon this feedback model to map the trajectories of individual students as they are involved in feedback behavior during a problem-solving task.

We use ProM (Günther & Van Der Aalst, 2007), an open-source process mining tool to analyze and visualize the process of interaction between individual student's feedback-seeking questions and responses. It uses multiple algorithms to shed light on the sequence in which a process occurs. Here, we use the Directly Follows Graph, Extended Causal Graph Miner algorithm, and Heuristics Miner algorithm (Weijters, Van Der Aalst, & De Medeiros, 2006) to visualize the trajectories of interaction among different feedback behaviors.

2. Research Questions

This study addresses the following research questions-

- With no specific instructions on collaboration, do students engage in feedback behavior during a problem-solving activity?
If yes, could we map the sequence of flow in which feedback behavior is manifested?
What does the sequence reveal about the trajectory of individual students’ feedback behavior in a problem-solving activity?

3. Methods

3.1 Participants, Data Sources & Task

Three male post-graduate chemistry students, who had completed an introductory course on stereochemistry, volunteered for a study based on tasks about representations of a medicinal drug. Students were seated at a round table and were individually given task-specific worksheets. They were free to work individually or collaboratively and the activity was video-recorded with minimal presence of the researcher. The video data formed the primary data source for this study. The video (1.5 hours) was transcribed using a combination of ‘Otter’, an online transcribing interface, and manual transcription. The study was approved by the IRB of the institute and students’ consent were received before the conduct of the study.

3.2 Data Coding

Following Hattie & Timperley (2007), we used codes to map feedback questions and responses pertaining to: a) form and structure of tasks (FQT & FRT: Feedback-seeking Question about the Task and Feedback Response on the Task), b) processes involved (FQP & FRP: Feedback-seeking Question about the Processing and Feedback Response about the Processing of the task), and c) monitoring (FQR & FRR: Feedback-seeking Question about Regulation and Feedback Response about Regulation of the task). We use codes S1, S2, & S3 for the three students.

4. Analysis and Findings

4.1 Students engage in feedback behavior on their own and PRoM algorithms mapped the sequence of the flow of feedback questions and responses

Figure 1 a) A feedback-seeking episode (Adapted from Swamy et al., 2022), b) a ‘Directly Follows Graph’ (DFG) representing the path for all the participants, (c) an Extended Causal Activity Graph, d) Heuristic Net Graphs for each student

Figure 1 a) depicts two students engaging in feedback behavior during the activity even in the absence of no specific instructions on collaboration. Further, a Directly Follows Graph (‘DFG’; Figure 1 b)) and an Extended Causal Graph (Figure 1 c)) mapped the overall flow of feedback events moving from the start (green circle in Figure 1 b) & ‘start’ box in 1 c) to the end (red circle in 1 b) & ‘end’ box in 1 c) of the problem-solving task for the participants. The nodes and the lines in the two graphs represent feedback behavior (questions or responses) and the transitions in these behaviors respectively. The numbers within the DFG nodes give the count of the events. Nodes with the greatest number of events are depicted in darker shades and vice-versa. The graphs reveal that the maximum number of feedback-seeking questions were asked about the task (FQT-21) and the maximum number of feedback responses were given
on the process (FRP-52), suggestive of a potential emphasis on process strategies to tackle questions on task. Further, from both graphs, we find that FQT (21 events) & FQP (10 events) behavior dominate the beginning phase and FRP (52) & FRR (16) behavior dominate the end phase of the problem-solving activity, suggesting that

4.2 Individual differences in interaction trajectories

Heuristic Net Graphs (Fig. 1 d)) mapped the trajectory of all students’ interaction between feedback questions and responses from the start (green circle) to the end (orange circle) of the problem-solving task. We find that S3 had only 34 events of feedback exchange, while S1 & S2 had more than 60 events for the process. We also find that S3 asked extremely few questions (7) overall. Thus, we can conclude that S1 and S2 were more actively involved in feedback behavior than S1. Thus, the graph gives an insight into the dynamics of interaction among the participants during the problem-solving process.

5. Discussion

In summary, we found that the participants naturally drifted toward collaborative problem-solving. The use of PRoM tools revealed that feedback-seeking questions on ‘task’ and ‘process’ dominated the start of the activity, while feedback responses on ‘process’ and ‘regulation’ dominated the end of the activity. Based on this, we suggest designing well-thought-out focus questions at the beginning of problem-solving exercises to both scaffold and facilitate students’ learning. The Heuristics Net graph revealed answers to broader questions like- ‘Who was more/least engaged in the feedback event?’ to finer questions like- ‘What form of feedback event was prominent in an individual’s interaction with the other group members?’ Our work contributes to the field by visualizing the process of feedback behavior and mapping the sequence of feedback events at both individual and aggregate levels. It has implications for both pedagogy and research, where such insights could be used to design scaffolds to promote scientific discourse (Seng & Hill, 2014), infer design principles for facilitating collaborations, and/or use/design tools to better visualize problem-solving behavior.

Acknowledgements

We would like to thank all the participants of the study. We also extend our gratitude to Narasimha Swamy for his help in providing access to the data.

References


Rattan, Savani, Chugh, & Dweck (2015). Leveraging mindsets to promote academic achievement: Policy recommendations. Perspectives on Psychological Science, 10(6), 721-726.


ICCE 2023 Learning Outcomes of Computer Programming and Information Technology -
Integrated Courses for Non-Computer Science Majors: Case Study of a Public Research University in Taiwan

Che-Yu Hsu*, Feng-Nan Hwang†, Tseng-Yi Chen*, Chia-Hui Chang

*Office of Institutional Research, National Central University, Taiwan
†Department of Mathematics, National Central University, Taiwan

Abstract: This study investigates non-CS major students’ performance in computer programming and Information Technology-integrated (ITI) courses at a Taiwanese research university. Non-CS students struggle in introductory programming due to syntax and logical thinking limitations, resulting in lower grades compared to advanced programming. Similarly, ITI course grades are lower due to subject-specific demands. Gender and entry channels impact outcomes, with females excelling in informal learning. Favorable results are seen in individual applications and Multi-star Projects. Challenges include different learning paces and cultural adjustments. Regression analysis shows Introduction to Computer Programming (ICP), Advanced Computer Programming (ACP) and gender significantly affect ITI performance, explaining 24% of its variance. Recommendations include diverse teaching methods, problem-solving guidance, practical programming, collaboration, and project participation to enhance skills.

Keywords: Computer programming, Information technology-integrated courses, learning outcomes, entrance channels, longitudinal study

1. Introduction and Key Questions

The 2016-2020 Information Education Blueprint aimed to enhance deep learning and digital citizenship through IT tools, aligning with Taiwan’s emphasis on computational thinking for problem-solving (Kalelioglu et al., 2016). The "Higher education sprout project” achieved 60% participation by June 2020, indicating successful non-CS programming course implementation (Ministry of Education, 2020). However, non-CS students face challenges in programming due to logic, syntax, and abstract thinking (Weng et al., 2014). Therefore, the research questions for this study are as follows:

- Are there grade differences among non-CS students in ITI courses, including programming skills in other subjects?
- Do gender and entrance channels influence the learning outcomes of non-CS students in ITI and programming courses?
- Is it possible to predict non-CS students’ performance in ITI courses using their programming course grades and relevant background factors?

2. Methodology

2.1 Participants and Computer Programming Implementation Courses Information
The participants in this study are non-information field professionals who have successfully completed three types of courses: ICP, ACP, and ITI courses. The total number of participants is 367. Among these students, 69% were male and 31% were female. Entrance channels included individual applications (39%), Multi-star projects (15%), exams (35%), and special programs/transfer/international admissions (11%). This reflects the Ministry of Education’s admission policies and highlights effective recruitment and selection strategies.

According to the curriculum design of the individual school, the course is divided into the following three parts:

1. ICP Course: Students will learn basic concepts of computer programming, including syntax, fundamental logic, and programming skills.
2. ACP Course: Students will delve into more complex programming techniques and strive to achieve specific programming goals.
3. ITI Course: The ITI course integrates subject knowledge with program analysis, including topics such as statistics. This will enable students to apply computer programming skills to specific disciplinary areas and solve relevant problems.

Grades earned by students may be obtained repeatedly in these classes, and these grades will be represented through the calculation of average semester grades. Performance differences were analyzed using weighted adjustments: ICP (15%), ACP (25%), and ITI (60%) of Computer and Technology (CT) Grades.

2.2 Statistical Analysis Tools

For data analysis, descriptive statistics compare grade differences. An independent t-test assesses gender disparities in CT, followed by mean comparisons. One-way ANOVA explores diverse non-CS major backgrounds, with post hoc tests. ICP and ACP predict ITI grades. JASP 0.16 conducts all analyses.

3. Results

3.1 Analysis of Computer Programming and ITI Course Performance of non-CS Majors.

Table 1 displays a performance comparison of Non-CS students based on grades in ICP, ACP, and ITI courses using repeated measures ANOVA (p<0.05 or F-value=14.27). Results highlight a significant performance difference among these course categories. Post hoc analysis indicates superior performance in ACP compared to ICP and ITI courses. Meanwhile, ICP and ITI course grades were comparable, suggesting support is needed for students in comprehending unfamiliar subjects within their respective disciplines.

Table 1. Comparison of course grade and background variables

<table>
<thead>
<tr>
<th>Courses</th>
<th>F</th>
<th>P</th>
<th>Post hoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICP &amp; ITI</td>
<td>14.27</td>
<td>&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>Individual Application Channel</td>
<td>12.03</td>
<td>&lt;0.01</td>
<td>Multi-star Project &gt;Exam-based channel &amp; Oversea&amp; Transfer</td>
</tr>
</tbody>
</table>

3.2 To compare the learning effectiveness of the CT Grades among students with different background variables.

Historically, there existed a perception of gender-based inclinations toward literature and art, posing challenges in STEM fields for females (Wrigley-Asante, C et al., 2023). Gender’s impact on CT grades was examined using a pair sample t-test (t-value=0.001<0.05), The results demonstrated that female non-CS students outperformed males, as presented in Table 1.
Entrance channels' influence on academic performance was investigated through one-way ANOVA (F value=12.03, p<0.05). Post-hoc analysis revealed higher grades for individual applications and "multi-star project" students compared to overseas and transfer students. "Multi-star project" entrants also outperformed exam-based entrants, as presented in Table 1. Notably, individual application and "multi-star project" students demonstrated stable IT and academic performance, while overseas and transfer students made additional efforts to reach similar levels.

3.3 The Relationship Between Student Variables and ITI Grade

Multiple regression analyses showed that ICP grades, ACP grades, and gender collectively explain 24% of the variance in ITI grades, which is statistically significant. Post hoc comparisons revealed significant contributions from ICP grades (t=6.03, p<0.001), ACP grades (t=4.16, p<0.001), and gender (t=-3.1, p<0.001) to the explanatory power. ICP grades had the highest influence (β=0.32), followed by ACP (β=0.2), and then gender (β=0.15). Entrance channel didn't significantly impact the model. Multicollinearity was assessed using VIF, with values well below 5 (ICP: VIF=1.21, ACP: VIF=1.21, Gender: VIF=1.02), indicating no multicollinearity issues, as presented in Table 2.

The regression equation is: ITI grade = 48.59 + (0.23) * ICP grade + (0.17) * ACP + (-3.53) * Gender. Gender is a dummy variable; when 0, it represents female students' influence, and when 1, it represents male students. Female students outperform male students by an average of 3.53 points in ITI grades.

Table 2. Linear Regression of ITI Grade

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>Std err</th>
<th>T</th>
<th>p-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const</td>
<td>48.59</td>
<td>3.87</td>
<td>12.54</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>ICP</td>
<td>0.23</td>
<td>0.04</td>
<td>6.03</td>
<td>&lt;0.001</td>
<td>1.21</td>
</tr>
<tr>
<td>ACP</td>
<td>0.17</td>
<td>0.04</td>
<td>4.16</td>
<td>&lt;0.001</td>
<td>1.21</td>
</tr>
<tr>
<td>Gender (male)</td>
<td>-3.53</td>
<td>1.11</td>
<td>-3.1</td>
<td>0.002</td>
<td>1.02</td>
</tr>
<tr>
<td>Entrance channels</td>
<td>0.02</td>
<td>0.41</td>
<td>0.05</td>
<td>0.96</td>
<td>1.06</td>
</tr>
</tbody>
</table>

4. Conclusion and Recommendations

Non-CS students struggle in ICP and ITI courses compared to those skilled in ACP due to unfamiliar syntax and initial lack of logical thinking training. Females excel, challenging stereotypes. Individual applications yield better outcomes than transfers or overseas recruitment. Multi-star Projects outperform exams and overseas recruitment, showcasing diverse experiences. Regression analysis explains 24% of ITI grade variance; ICP grade, ACP grade, and gender influence ITI grade. Recommendations encompass diverse teaching methods, promoting practical programming, fostering collaboration, and encouraging participation in information competitions. Interactive learning environments and providing support for overseas students are essential to enhance the learning experience.

References


Ministry of Education. (2020). 60% of college students have taken courses related to programming, the Ministry of Education's "Promotion of University Programming Teaching Program" has seen results. Electronic newspaper of Ministry of Education. https://epaper.edu.tw/print.aspx?print_type=topical&print_sn=1154&print_num=928


DLOT: An open-source application to assist human observers

Ashwin T Sa*, Danish Shafi SHAIKH, & Ramkumar RAJENDRAN

a Indian Institute of Technology, Bombay
*ashwindixit9@gmail.com

Abstract: Adaptive intelligent educational systems are gaining popularity, offering personalized learning experiences to students based on their individual needs and styles. One crucial feature of such systems is real-time personalized feedback. However, identifying real-time learning processes impacting student performance remains challenging due to data volume constraints. Current research often relies on labor-intensive human observation, which is time-consuming and not scalable. To efficiently collect real-time data, an observation tool is essential. Qualitative/Mixed Method research explores participant experiences in education, social science, and healthcare, utilizing methods like focus groups and observations. However, these methods can be labor-intensive, particularly in maintaining observation time intervals. Existing tools lack comprehensive support for education-focused focus groups and observations. To address these issues, this paper introduces the Data Logging and Organizational Tool (DLOT), a flexible tool designed for qualitative studies with human observers. DLOT offers customizable time intervals, cross-platform compatibility, and data saving and sharing options. The tool empowers observers to log timestamped data and is available on GitHub. The DLOT was validated through two studies. The first study predicted students' affective states using real-time annotations collected via DLOT, observing 30 students in each class. The second study created multimodal datasets in a computer-enabled learning environment, observing 38 students individually. A successful usability test was conducted, offering a potential solution to challenges in real-time learning process identification and labor-intensive qualitative research observation.

Keywords: Human Observation, Annotation Tool, Application, Data Logging, Computer Assisted Direct Observation, Minimal Attention User Interfaces

1. Introduction

Adaptive intelligent educational systems are becoming increasingly popular, particularly in distance learning and online education. These systems are designed to provide personalized learning experiences to students, catering to their individual learning needs and styles. One of the key features of such systems is the provision of personalized feedback to students based on their learning processes. To be effective, this feedback needs to be provided in real-time as the student interacts with the learning environment (Mousavinasab et al., 2021, Ashwin & Guddeti, 2020, Taub et al. 2021).

Identifying the various learning processes that contribute to a student's real-time performance is a challenging task. These processes include self-regulated learning, cognitive engagement, metacognition, affective states, behavioral and interactional dynamics with peers, collaborative learning, uncertainty, feedback seeking, empathy, and group dynamics. To train machine learning models to detect these processes in real-time, a large amount of annotated data is necessary to identify, classify, or predict these constructs (Sinatra, Heddy, & Lombardi, 2015, Praharaj et al. 2021, Ashwin & Guddeti, 2019, Munshi et al. 2018).

Current research in this area involves human observation, where video recordings are analyzed after the study has concluded. However, this method is time-consuming, labor-intensive, and not scalable. Automated machine learning models also face challenges in real-time identification of these constructs due to insufficient data (Wu et al. 2022, TS & Guddeti, 2020). Hence, there is a need for a reliable observation tool that can collect data quickly and efficiently in real-time. Such a tool could revolutionize the field of adaptive intelligent
educational systems, providing personalized feedback to students based on their real-time learning processes, leading to improved learning outcomes.

On the other hand, Qualitative/Mixed Method research is an approach to research that focuses on exploring the experiences, attitudes, and perspectives of participants in a study. It is often used in education, social science, and healthcare research, where understanding human behavior is a key objective. The research involves collecting data through a variety of methods, such as questionnaires, interviews, focus groups, tests, observations, and secondary data (Johnson, Burke, & Turner, 2003). While some methods like questionnaires and tests, can be conducted without human observation, others, such as focus groups and observations, require human observers to gather data. These methods are crucial in providing insights into human behavior, emotions, and experiences. However, the process of human observation can be labor-intensive, especially when there are a large number of subjects or the observation period is lengthy. In addition, maintaining the exact time interval between observations can be challenging for human observers. An application or tool that automatically sets the timer and provides a prompt for the observer to log the data would be incredibly useful in such cases.

To facilitate the collection of data with specific labels, we have developed this app, drawing inspiration from the Human Affect Recording Tool (HART). Similar to various tools commonly employed in qualitative and mixed-method research, such as Teamscope, Open Data Kit (ODK), and KoboToolbox, which find utility in questionnaires, tests, and secondary data collection, our app serves a comparable purpose. It's worth noting that while these established tools are often used for activities like questionnaires, tests, and secondary data collection, they may not be ideally suited for tasks like focus groups and observations. Additionally, the adaptation of new technology might involve a learning curve for both participants and researchers, potentially influencing the efficiency and accuracy of the data collection process. In the context of educational research, the Human Affect Recording Tool (HART) has been utilized for Quantitative Field Observations (QFOs), primarily focusing on monitoring students' affect and behavior (Ocumpaugh et al. 2015). However, in response to this limitation, our innovation aims to empower human observers by providing a mechanism to log data along with timestamps.

Hence, this paper contributes to a DLOT (Data Logging and Organizational Tool) developed to aid qualitative and also quantitative studies involving human observers. DLOT is flexible and can be customized to set time intervals, works with different operating systems, works for the individual, team (or group), different sets of codes/labels, and offers various options for saving and sharing data.

Moreover, it's pertinent to emphasize that the central focus of this article is to present a concise and comprehensive overview of the Data Logging and Organizational Tool (DLOT). Rather than a typical research study, the article is tailored to provide an in-depth description of DLOT's features and functionality. The rest of the paper is organized as follows: Section 2 provides details of DLOT, Section 3 discusses the validation, and Section 4 concludes the paper with future directions.

2. DLOT
The DLOT tool was developed using Figma and React Native. Data export was handled using the docx and xlsx packages. State management was implemented using useState hooks, and dynamic runtime changes were handled with the useEffect hook. The app development and testing were done using Expo. Additional details about the DLOT tool development are mentioned in Appendix A1. The complete details, including the readme information, can be found in the GitHub repository.

The DLOT tool offers a wide range of flexible options that cater to various research requirements:

Class Labels or Categories: Researchers using DLOT can create custom labels or categories tailored to their study objectives. For example, Figure 1a illustrates class labels

1 https://github.com/danishsshaikh/DLOT/Appendix A
such as "engagement" to "on-task," which can be modified to suit different studies. For instance, in a metacognition study, labels can be defined as "planning," "monitoring," "evaluating," and more, based on the specific research context. Prior to using the DLOT tool, researchers need to establish a set of categories or class labels that will be used throughout the study. This step involves defining the specific labels that will be assigned to observations. Once set, these categories remain consistent and cannot be changed during the course of the study.

**Time Flexibility:** DLOT offers researchers the flexibility to set the observation timer according to their study requirements. By default, the timer is set to 10 seconds, providing a suitable time frame for observations. The timer duration is conveniently displayed atop Figure 1a.

**Selection of Categories:** DLOT empowers researchers to design observation interfaces that include both radio buttons and checklists, as demonstrated in Figure 1a. This feature enables researchers to make multiple selections based on their study’s specific needs, enhancing the richness and versatility of data collection.

![Figure 1.](image.png)

Figure 1. There are three screenshots attached that illustrate the data logging and organizational tool. The first screenshot (Figure a) displays the interface where the student's name, class labels, timer, and a log button can be found. Once the logging process is completed, the data can be downloaded in either xlsx or docx format. After selecting the desired format, the data can be stored locally on the mobile device or in any other preferred location, as depicted in Figure b). In Figure c), a sample of the logged data is displayed in a particular format. In this format, the student's name or ID acts as the primary key, and the data is organized based on the timestamp.

**Adjusting Observations (Number of Students and Group Size):** DLOT allows researchers to customize the number of students observed, accommodating studies involving individuals, small teams, or larger groups like classrooms. Researchers can adapt the tool to their specific research setting.

**Platform Compatibility and Data Storage:** DLOT is compatible with both Android and iOS devices, ensuring broad accessibility for researchers using different operating systems. Additionally, the tool supports multiple data storage formats, including text and xlsx. This versatility facilitates easier data analysis and sharing, streamlining the research workflow.
**File Sharing Options:** Once data logging is complete, users can choose to store the final data locally or share it through various options such as Google Drive, WhatsApp, Telegram, Bluetooth, LINE, etc. Figure 1b illustrates the convenient file sharing functionality of DLOT.

**Open-Source Availability:** DLOT is available as an open-source application with its GitHub code and official webpage guidelines and read-me information. Sample screenshots of DLOT are shown in Figure 1. The tool also includes a discussion forum on GitHub where researchers can ask questions about using the app or modifying default values, among other topics.

**Training Observers for Data Annotation:** To utilize the DLOT tool effectively, human observers need to undergo training in annotating or classifying data based on standard definitions or study guidelines. For instance, in a basic emotion detection study, multiple observers should observe the same student(s) and provide labels following established guidelines, such as a facial action coding system. Observers can choose various annotation methods, including annotating together or independently and later discussing any disagreements. Importantly, observers must adhere to the specific purpose of the observation. For example, suppose the annotations aim to train a classification model like convolutional neural networks. In that case, guidelines should emphasize labeling based on the specific instance observed rather than relying on prior frames or external sources of knowledge.

The following steps should be followed to use the DLOT:
1. Train the human observers on the required class labels or categories and establish inter-rater reliability.
2. Install the application.
3. Check if the required parameters are present in the application.
4. If yes, go to step 5; Else, modify the open-source code according to the requirements.
5. Set the parameters and collect data.
6. Save the data. (More details are in Appendix B)

By embracing flexibility, customization, and compatibility, DLOT empowers researchers to efficiently collect and analyze qualitative data in a manner that suits their specific study objectives and research settings.

### 3. Validation

The Data Logging and Organizational Tool (DLOT) was validated using two studies. The first study focused on affective state prediction in a classroom setting, where real-time annotations were collected using DLOT. A total of 30 students were observed in each class, and their affective states, including engaged, boredom, confusion, frustration, and neutral, were predicted. The timer interval was set to 5 seconds as per the study requirement. The second study incorporated DLOT as a component in generating multimodal datasets (interaction logs, screen recordings, human observations and facial expressions), which involved 38 participants in a computer-enabled learning environment. This study involved observing a single student at a time, with a time interval of 10 seconds throughout the study. Three observers used the tool in the first study, while five used it in the second. The validation process encompassed an evaluation of system usability, conducted via the System Usability Scale (SUS) analysis, yielding an average score of 93. SUS, a widely utilized questionnaire, employs a 5-point Likert scale and subsequently transforms the scores to a range of 0 to 100 (Brooke, 1995). The attained score of 93 reflects the collective assessment by participants and underscores the commendable usability of the system. Observers reported that the tool was handy and saved them significant time. The user interface was deemed easy to navigate, and the installation process was straightforward.

DLOT was tested on different operating systems, including Android and iOS, and performed well without technical glitches or issues. The open-source code of DLOT was made

---

2 https://github.com/danishsshaikh/DLOT
3 https://danishsshaikh.github.io/DLOT/
4 https://github.com/danishsshaikh/DLOT/Appendix B
available for observers. One observer from the study one used it and found it easy to modify and make changes. He also mentioned the information provided in the open-source code was clear and concise, and the accompanying discussion forum proved helpful in addressing queries and concerns.

During the studies, DLOT was used continuously for extended periods while observing 30 students, and no technical issues such as hanging or data loss were experienced. Overall, the validation process and user feedback demonstrate the reliability, usability, and effectiveness of the DLOT tool in diverse educational contexts.

Since there are not many manual note-taking tools in the education domain, it is relevant to compare DLOT with HART, the only existing tool that is similar to DLOT in terms of observing students and noting down their states with a time stamp. The comparison and advantages of DLOT over HART are as follows: Fixed Class Labels: HART has predefined class labels or categories based on Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) (Ocumpaugh, 2015), limiting the customization options for researchers. Time Limitations: HART imposes a fixed countdown clock of 3 minutes for each observation, which restricts the coding time to 20 seconds for BROMP observers. Single Selection: HART only allows the selection of one class label or category per student within a time interval, using radio buttons. Platform Compatibility: HART is designed exclusively for the Android platform, limiting its accessibility to researchers. Data Storage Formats: HART stores data only in the txt format, which may limit the options for data analysis and sharing. File Sharing Options: HART data files are stored locally on Android devices and can be shared via email or transferred to computers using USB cables. As explained in the previous section, DLOT offers advantages compared to HART in terms of flexibility in labeling, time management, selection options, platform compatibility, data storage formats, and file sharing options.

Limitations: In this study, there are several limitations of the Data Logging and Organizational Tool (DLOT). While DLOT demonstrates potential for real-time qualitative studies, a comparative usability analysis with existing tools was not conducted. Additionally, the presence of observers using DLOT could introduce bias, demanding caution. Security and privacy concerns inherent to digital platforms must also be considered. Furthermore, DLOT’s reliance on smartphone technology and internet access might restrict its use in certain contexts.

In addition to its use in the education sector, this tool has the potential to be utilized in multiple other domains. For instance, in research studies, researchers can use the app to log data regarding participant behaviors, reactions to stimuli, or experiment results. Clinical psychologists can also use the app to record data concerning patient sessions, including symptoms, behaviors, or progress over time. Moreover, coaches or athletes can utilize the tool to log data related to training sessions or competitions, such as times, distances, or personal bests.

4. Conclusion

The Data Logging and Organizational Tool (DLOT) is a flexible and user-friendly application developed for data collection and analysis in educational research. It offers researchers the ability to create custom class labels or categories tailored to their study objectives, adjust observation timers, and choose from multiple selection options for data collection. DLOT is compatible with Android and iOS platforms, ensuring accessibility for researchers using different operating systems. It supports various data storage formats, facilitating easy analysis and sharing of data. The tool also provides convenient file sharing options, allowing users to store data locally or share it through different platforms. Being an open-source application, DLOT empowers researchers to modify and customize the tool according to their specific needs. The GitHub repository provides comprehensive information and a discussion forum for user support. Validation studies conducted on affective state prediction and multimodal dataset creation have shown the reliability and usability of DLOT. Observers reported that the tool was highly useful and time-saving, with an intuitive interface and straightforward installation process. DLOT performed well on different operating systems, ensuring its effectiveness across platforms.
Several future directions for the DLOT include adding data visualization capabilities, user authentication, cloud storage integration, customization in different sets of options, integration with other educational apps, and audio and video capabilities for recording and analyzing speech for unstructured data.

References


Supporting Learning Through Affordance-Based Design: A Comparative Analysis of "BioVARse" and a Standard Textbook Companion Application in Biology Education

Devanshu SAINDANEa, Sunny Prakash PRAJAPATIa & Syaamantak DASA*
aIDP in Educational Technology, Indian Institute of Technology Bombay, India
*syaamantak.das@iitb.ac.in

Abstract: Textbooks as an educational artefact has evolved immensely with technology, especially through mobile applications. When used in conjunction with textbooks, mobile applications have the potential to provide a unique set of affordances to support learning. This paper presents the design of “BioVARse”, a textbook companion app for Biology subjects that provides these affordances to the learners. A comparative analysis of these learning affordances with a standard textbook companion application is presented, and a pilot study for assessing the usability is also performed. The preliminary results indicate a significantly higher usability of BioVARse, when compared to the standard application. Additionally, the paper also discusses the possibility of improving the learning experience through an affordance-based design of a textbook companion app.

Keywords: Multimedia Educational Content, Textbook companion application, Augmented Reality, Affordance-based Design, Usability

1. Introduction and Background

Textbooks are the primary learning materials that are used for studying a subject by learners. Appropriate technological enrichment can provide access to static and dynamic content, such as in Phygital textbooks (Prajapati & Das, 2023), which can impact learning outcomes of an individual, such as their reasoning behaviour (Karnam et al., 2020). Currently, textbooks are enriched with dynamic content through the quick response (QR) codes or through Augmented Reality (AR) markers on the printed textbook page (Danaei et al., 2020; Ghaem Sigarchian et al., 2018). This content is then accessed through digital devices, such as mobile phones, which can be termed as the textbook companion applications.

In Science, topics in Biology (and similar subjects) require visualisations for learning of various concepts, as inappropriate visualisations may result in development of misconceptions (Yusof et al., 2020). AR can be used in this situation to effectively facilitate immersive visualisations using dynamic content, and it can be provided to learners through a mobile application that can be used in different learning scenarios (Susilo et al., 2021). Companion apps that are available to the learners primarily focus on the content accessibility, failing to address the issue of their integration into learning activities.

Consideration of affordances that are crucial in learning can be an important aspect that should be considered for such applications (Hammond, 2010). In human-computer interaction, "affordance" refers to the perceived or actual action possibilities an object offers a user. Though lacking a universal definition, this work defines it as the set of interactions facilitated by a tool that supports learning. This study details the design of a textbook companion app (BioVARse), based on affordance analysis and comparison with a standard app, and elucidates a usability study of these application.
2. Design Space for BioVARse

A markerless AR app, 'BioVARse,' was developed using Unity 3D, for 10th and 11th-grade students studying Biology. It covers five NCERT curriculum topics, with textbook diagrams coded as markers for 3D models. These topics were: "Body fluids and circulation", "Breathing and exchange of gases", "Digestion and absorption", "Excretory products and their elimination", and "Structure of human eye". The name BioVARse combines concepts of metaverse, virtual reality, and augmented reality. The system architecture, detailed in Figure 1, elucidates the integration of AR, Quiz, and App modules. The prototype is accessible on Github1, and the topics were selected from federal textbooks based on open-source AR content availability, and resources availability on DIKSHA. The key affordances of BioVARse applications specifically designed to support learning of the topic are: Note-taking, Voice-over narration of the text, Access to 3D AR content, Formative assessment for the topic, Access to static content, and Sharing of information. The affordances due to mobile devices were same for both applications, i.e., DIKSHA and BioVARse, which are the ability to interact with learning content through touch, scroll, text input, etc. (Berthelsen & Tannert, 2020). The affordances (listed above) were facilitated to the user through three primary components of the app, which are: a) Learn (for static and dynamic content accessibility), b) Explore (for engagement with content), and c) Test (for formative assessments and feedbacks). These affordances are elaborated in Figure 2 below, which were used as a comparison metric for both the applications.

Figure 1. System architecture of BioVARse application.

3. Study Design and Results

The usability of the BioVARse app was assessed using the System Usability Scale (SUS), chosen for its reliability and ease of use. Eleven participants (chosen through convenient sampling) interacted with the app on the topic 'Human Heart (Body fluids and circulation),' and their SUS survey responses were recorded online. Instructions were provided, and users were encouraged to engage with AR models, multimedia content, and assessments. The SUS survey was administered for both the BioVARse and NCERT DIKSHA apps, and usability scores were compared using the Mann-Whitney U test, revealing a significant difference (U = 13.5, p < 0.005). The results indicate acceptable usability for BioVARse and poor usability for the standard DIKSHA application.

---

1 BioVARse: https://github.com/SPARTA-Research-Group-ET-IITB/BioVARse
### Figure 2: Key learning affordances and how they are facilitated in the two mobile applications

<table>
<thead>
<tr>
<th>Access to 3D AR content: 3D models that can be manipulated through hand gestures can provide enhanced spatial perception of the object, which is mentioned in the text (Swamy K L et al., 2018).</th>
<th>Available</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Formative assessment: Operationalised in the form of a set of multiple choice questions with four options, which can be mediated by an instructor in the classroom or could be used independently by learners with the primary goal of providing feedback on the learning of the content (Bennett, 2011).</td>
<td>Facilitated through MCQs after every topic</td>
<td>Facilitated through MCQs after every topic</td>
</tr>
<tr>
<td>5. Access to text and images: interaction with texts and images in the textbooks through viewing on mobile screen, navigating across and within textbook page, resizing, etc are also afforded, similar to other digital learning materials (Berthelsen &amp; Tannert, 2020).</td>
<td>Available</td>
<td>Available</td>
</tr>
<tr>
<td>6. Sharing of personalised information: personalised information such as screen recordings, and notes can be shared by the students with their peers, hence, providing ways to collaborate in and outside the classroom environment.</td>
<td>Personalised information (such as notes, recordings, etc.) can be shared by user to their peers</td>
<td>Only the standard content can be shared by the user</td>
</tr>
<tr>
<td>7. General affordances</td>
<td>Available</td>
<td>Available</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>References</th>
</tr>
</thead>
</table>
Investigating Programming Performance Predictability from Embedding Vectors of Coding Behaviors

Ikkei IGAWA*, Yuta TANIGUCHI*, Tsubasa MINEMATSU*, Fumiya OKUBO* & Atsushi SHIMADA*
*Kyushu University, Japan
* igawa@limu.ait.kyushu-u.ac.jp

Abstract: Understanding students’ coding behaviors is crucial for providing targeted support in programming education. Automatic analysis of coding behaviors using machines can address the limitations of manual monitoring. Previous studies focused on coding behavior representations without considering differences relative to a model answer. We propose embedding vectors that capture these differences, enabling the distinction between simple and complex code solutions. Evaluating these vectors by predicting assignment scores, we achieved over 15% higher accuracy compared to conventional methods. This approach has the potential to enhance teachers’ understanding of students’ coding behaviors and improve support in programming education.

Keywords: Programming education, educational data analysis, CodeBERT, programming performance prediction

1. Introduction

Understanding student coding behaviors in programming courses offers several advantages, including identifying struggling students and pinpointing areas of difficulty. This enables teachers to provide targeted support and effective feedback, enhancing students’ understanding of programming (Leinonen et al., 2019). However, directly monitoring all students’ programming activities is not feasible for teachers due to large class sizes and limited resources. Therefore, the use of machine-based automatic analysis of student coding behaviors becomes essential.

Previous studies have conducted the automation of programming log analysis (Mao et al., 2021). However, many of them do not consider the difference between student codes and a model answer when calculating coding behavior representations. By considering these differences, it becomes feasible to distinguish between students who have correctly solved a problem with simpler code closer to the model answer and those who have written more complex code. To address this, we propose the generation of coding behavior representations that consider the deviation from the model answer. Our objective is to utilize these representations, derived from the coding activity logs of students during programming assignments, to predict their performance accurately. By achieving accurate predictions of assignment scores based on coding behavior representations, we demonstrate that these representations effectively capture factors that affects their programming performances.

2. Proposal Method

To create embedding vectors that capture students' coding behaviors, including their ability to write code successfully and smoothly, as well as their tendency to explore different approaches through trial and error, we propose some embedding methods.
First, we utilize the pre-trained CodeBERT model (Feng et al., 2020) without fine-tuning to convert students’ code into embeddings. CodeBERT is a pre-trained bimodal model for PLP with Transformer-based neural architecture. Next, to compute an embedding vector for each student’s coding behavior, we utilize the answer code provided for a given programming assignment. By considering the answer code as a reference, we quantify the characteristics of a student’s coding behavior relative to this reference code. To accomplish this, we introduce the concept of answer-directed vectors (ADVs), which are vectors pointing from each code written by a student to the answer code.

After calculating the ADV for each student’s code, we generate two vectors: the mean of the ADVs (ADV Mean) and the sum of the ADVs (ADV Sum). We expect that struggling students will have a larger ADV Sum due to their trial-and-error approach and significant code differences. The ADV Sum can be calculated like below:

$$\sum_{i=1}^{n} (e_{\text{Answer Code}} - e_{\text{Code}}^i),$$

Whereas $e_{\text{Answer Code}}$ indicates the embedding vector of the answer code, and $e_{\text{Code}}^i$ indicates the embedding vector of the $i$-th code written by a student to solve the assignment. Therefore, we can calculate the ADV Mean by dividing ADV Sum by $n$.

In this study, we evaluate the ADV Mean and ADV Sum as representative vectors for student coding behaviors, examining their effectiveness in capturing and representing the nuances of students’ approaches to programming assignments.

3. Experiment

3.1 Experiment Settings

To assess the effectiveness of the generated vectors in representing students’ coding behaviors, we conducted an experiment to predict students’ programming assignment scores. We utilized WEVL (Taniguchi et al., 2022), an online coding application used in programming courses at our university, to analyze the programming logs collected during a Python programming exercise course in 2020. Throughout the course, students received weekly programming assignments to be completed within a week.

We categorized students into two groups based on their scores: “Comp.” (students who obtained full scores) and “Incomp.” (students who did not). Then, we tested the predictability of the vectors by predicting which students were in Incomp., using machine learning. For binary classification, we employed a Random Forest with a maximum depth of 5. We then performed a $k$-fold cross-validation with $k=3$ and calculated the average scores of the four types of prediction performance metrics: recall, precision, accuracy, and f-measure.

In this experiment, we examined the “concat” assignment given during the middle stages of the course. The number of programming logs was 6,301. Among the 63 students who engaged with the “concat” assignment, 32 students belonged to the Incomp. group, while the remaining 31 students belonged to the Comp. group.

3.2 Results

Table 1 shows the prediction results of each vector in a programming assignment given in the middle stage of the course. For comparison, we used edit distance and embedding distance to calculate how the student code was different from the answer code. First, we can see that the vector generated by summing up ADV exhibits the highest scores in all the performance metrics. Second, we can see that the sum method outperforms the mean method in prediction scores in almost all the metrics.
Table 1. Results of prediction on the concat assignment. The highest values for each performance metric are highlighted in bold. The chance rate of the prediction is 0.5079.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit Distance</td>
<td>Mean</td>
<td>0.4453</td>
<td>0.5039</td>
<td>0.4762</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>0.4913</td>
<td>0.5371</td>
<td>0.5397</td>
</tr>
<tr>
<td>Embedding Distance</td>
<td>Mean</td>
<td>0.5030</td>
<td>0.5571</td>
<td>0.5397</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td>0.5939</td>
<td>0.5563</td>
<td>0.5397</td>
</tr>
<tr>
<td>ADV</td>
<td>Mean</td>
<td>0.6212</td>
<td>0.6474</td>
<td>0.6349</td>
</tr>
<tr>
<td></td>
<td>Sum</td>
<td><strong>0.7698</strong></td>
<td><strong>0.7098</strong></td>
<td><strong>0.6984</strong></td>
</tr>
</tbody>
</table>

4. Discussion

Looking at the Table 1, we can see that the sum method consistently outperformed the mean method across almost all methods. Considering the reason why this happened, a possible explanation is that students who are less proficient in programming and struggle with coding may require repeated trial and error to arrive at a solution. This increase in the number of attempts results in a greater sum. If we consider distance or vector size as penalties, we can accumulate them by employing the sum method, enabling the identification of students struggling with coding. Indeed, both Edit Distance Sum and Embedding Distance Sum are proportional to the number of attempts, but their scores did not increase as much as ADV Sum. This indicates that the direction of the ADV, as well as its size, contributes to prediction. In summary, our findings suggest that the ADV Sum represents factors that significantly affect student performance.

5. Conclusions

This paper presented methods using CodeBERT to convert students' coding behaviors into 768-dimensional vectors. The ADV Sum method, incorporating differences between student and answer code. CodeBERT-based representations outperformed edit distance, providing valuable insights for educators to understand and support students' learning needs in programming education. Enhanced feedback based on these insights can improve student learning outcomes.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR22D1, JSPS KAKENHI Grant Number JP18H04125, JP21K17863 and JP22H00551, Japan.

References


Unveiling Learners’ Interaction Behavior in Virtual Reality Learning Environment

Antony PRAKASH*, Ramkumar RAJENDRAN
Interdisciplinary Programme in Educational Technology, Indian Institute of Technology Bombay, India
*antony80004@gmail.com

Abstract: VR is increasingly being utilized in various domains, including education, due to its unique characteristics. Research in this area often relies on physiological sensors, eye-trackers, VR device orientation, human observers, and pre-test and post-test to collect data for quantitative studies and on questionnaires, surveys, and interviews to collect data for qualitative studies in VR Learning. However, there is a dearth of reliable data sources for studying learner behavior in VRLE, and minimal efforts have been made to automatically collect behavioral data in this context. Furthermore, there is a lack of studies that investigate learning processes through the lens of learners’ dynamic interaction behavior in VRLE. To address these gaps, we have developed a real-time data collection mechanism that automatically logs learners’ interaction behavior in VRLE, including timestamps. This mechanism was deployed in a room-scale VRLE called MaroonVR, and a study was conducted involving undergraduate engineering students. The main objectives of this paper are to identify differences in interaction behavior between high and low performers and to develop an optimal predictor model to predict the learning outcome using learners’ interaction behavior in VRLE. Furthermore, we propose that the study's findings can be utilized to model learners' behavior in VR and to provide scaffolding and adaptive personalized VR learning content.

Keywords: Interaction Behavior, VR Learning Environment, Predictor Model, Adaptive Personalized Learning Content

1. Introduction

Virtual Reality (VR) immerses users in a simulated environment, allowing interaction with virtual objects and creating a realistic experience (Wade, Zhang, Bian, Fan, Swanson, Weitlauf, & Sarkar, 2016). Its distinctive qualities of immersion, interaction, and imagination have led to its application in various domains, including automotive, military, healthcare, sports, and education. In education, VR is utilized to teach concepts that are not visible to the naked eye, such as DNA strands (Sharma, Jin, Prabhakaran, & Gans, 2018) and the human circulatory system (Pathan, Rajendran, & Murthy, 2020). It also enables the exploration of inaccessible places like outer space and ancient civilizations. Additionally, VR provides a safe environment for experiencing hazardous scenarios, such as firefighting, welding, and oil refinery operations, which would be dangerous in real life. VR technology has transformed education by offering immersive and engaging learning experiences that enhance understanding and provide practical training opportunities in a wide range of subjects.

The qualities of VR have led to a significant increase in research examining its use in education. Most of the studies were conducted to measure the impact of VR on learning by collecting data from pre-test and post-test. Moreover, the educational researchers used the data collected from self-reported questionnaires, interviews, and surveys (Radianti, Majchrzak, Fromm, & Wohlgenannt, 2020) to measure the user experience, engagement, and usability of the VR systems and to compare VR-aided and VR-non-aided learning systems (Albus, Vogt, & Seufert, 2021). The researchers also used the data collected from the devices such as 1) physiological sensors to assess the affective state of the learners while performing the tasks (Feng, González, Amor, Lovreglio, & Cabrera-Guerrero, 2018), 2) eye trackers to assess the learners' intended area of interest (Wade, Zhang, Bian, Fan, Swanson, Weitlauf,
3) body trackers to adapt the size of the virtual objects with respect to the size of the users, and 4) orientation of the head-mounted displays (HMD) and handheld controllers (HHC) to assess the response time. In addition to the physical devices, human observers were also used to collect data in VRLEs related to the affective state (Feng, González, Amor, Lovreglio, & Cabrera-Guerrero, 2018) and the procedural performance (Santamaria-Bonfil, Ibáñez, Pérez-Ramírez, Arroyo-Figueroa, & Martínez-Álvarez, 2020) of the learners. However, the data provided by human observers are biased due to cognitive, social, and communicative causes. Moreover, the data provided by human observers also need to satisfy inter-rater reliability tests in order to become valid (Olmos-Raya, Ferreira-Cavalcanti, Contero, Castellanos, Giglioli, & Alcañiz, 2018).

All the existing studies reported that using VR technology in the education domain has resulted in a) better learning compared to other learning environments like simulation or Computer Based Learning Environments (CBLE) for procedure learning content (Santamaria-Bonfil, Ibáñez, Pérez-Ramírez, Arroyo-Figueroa, & Martínez-Álvarez, 2020), b) positive impacts such as improvement in learning gain, experience closer to reality, intrinsic motivation, level of interest, skills, and memory retention (Chavez, & Bayona, 2018). Although the existing studies measured the learning gain, the knowledge of how learners interact with VRLE and how they learn from the VR environment is not explored. That is, the existing studies have analyzed the impact of VR intervention on the learning outcomes, but the impact of interaction behavior in VRLE on the learning outcomes is still in its infancy. Hence, the interaction behavior of the learners leading to variation in the performance for different learners is not known. This is mainly due to the non-existence of a data collection mechanism that can log the learners' interaction behavior in VRLE. To address this gap, we developed a mechanism that is able to log all the interaction behavior of the learners in VRLE in real time along with the time stamp (Prakash, & Rajendran, 2022).

The interaction behavioral data collection mechanism we developed was deployed in MaroonVR (Pirker, Holly, Lesjak, Kopf, & Gütl, 2019), a virtual reality learning environment (VRLE) utilized for teaching the physics concept of electromagnetic induction. A study was conducted involving 14 undergraduate students from non-electrical engineering backgrounds, and their interaction behavioral data (IBD) was logged. The participants' interaction with the VRLE resulted in a positive learning gain (Prakash, Shaikh, & Rajendran, 2023). In this paper, we present the extraction of features, such as frequency and duration of action events, to evaluate their impact on the learning outcomes. Furthermore, we discuss the development of an optimal regression predictor model using the features extracted from the data logged in the IBD logger to predict the learning outcome. The predictor model thus developed can be used to early detect the performance of the learners. The knowledge of the early detection of the learners' performance can be used by the designers to design VR learning content that is able to provide the required scaffolding in the form of hints, and feedback to the learners in order to maximize the learning outcome.

The paper is structured as follows. Section 2 discusses the impacts, the learning outcomes, and data collected in the education domain using VR. The research questions addressed by this paper are also presented in section 2. The research methodology is briefed in Section 3 along with various analyses. The results of the analyses are presented in Section 4. The inferences made from the results are discussed in Section 5 along with the conclusion describing the limitations and the guidelines for future work.

2. Literature Review and Background

In this section, we first describe the works related to the impacts of VR on learning. Then we discuss the learning outcomes and the data collected to measure VR impacts and the learning outcomes in the existing studies. We also give a brief overview of the IBD logger and the data logged in it.
2.1 Impacts of VR on Learning

According to Radianti, Majchrzak, Fromm, and Wohlgenannt, (2020), the results obtained from employing immersive VR technology in various educational domains indicate an increase in engagement, time dedicated to learning tasks, and the development of cognitive, psychomotor, and affective skills. VR's ability to offer interaction, immersion (Hamilton, McKechnie, Edgerton, Wilson, 2021), and a first-person perspective (Mikropoulos, & Bellou, 2010) play a significant role in enhancing learning outcomes by providing realistic experiences, fostering intrinsic motivation, and increasing interest in learning. Chavez, and Bayona, (2018) asserted that no literature reported negative effects of using VR in the learning process. Nevertheless, the comprehensive literature review conducted by Hamilton, McKechnie, Edgerton, and Wilson in 2021 revealed a reduction in learning improvement with VR-based learning when contrasted with both desktop learning and conventional classroom learning. Despite the negative learning outcome observed in some studies, potentially attributed to the use of low-end mobile VR tools and the inclusion of factual learning content in VRLE, learners have demonstrated increased motivation and interest in learning when utilizing VR as opposed to traditional approaches and computer-based learning (Makransky, Terkildsen, & Mayer, 2019). The existing studies suggest that VR has a positive impact on learning, particularly in VRLE involving procedural learning content. However, there is a lack of literature that examines the impact of VR learning in relation to the dynamic behavior of the learners in the VRLE.

2.2 Learning Outcomes Measured in VR Learning Environment

Researchers have examined the effects of VR on learning by evaluating various learning outcomes, including cognitive, procedural, and affective skills (Hamilton, McKechnie, Edgerton, Wilson, 2021). Cognitive skills pertain to acquiring declarative knowledge, procedural skills involve psychomotor abilities, and affective skills are related to emotions and attitudes (Hamilton, McKechnie, Edgerton, Wilson, 2021). Among these skills, cognitive skills have received the most attention in VR studies. The evaluation of cognitive skills typically involves assessing knowledge acquisition, retention, and transfer. In existing studies conducted within VR learning environments (VRLE), the assessment of knowledge acquisition is typically done through pre-tests and post-tests, while knowledge retention is evaluated through delayed post-tests. Procedural skills, on the other hand, are assessed by measuring task completion time and the sequential order of accessing intermediate steps to accomplish a task (Feng, González, Amor, Lovreglio,& Cabrera-Guerrero, 2018). Affective skills are evaluated using questionnaires and physiological devices such as electrodermal activity sensors, photoplethysmography sensors, and multichannel physiological sensors (Radianti, Majchrzak, Fromm, & Wohlgenannt, 2020; Feng, González, Amor, Lovreglio,& Cabrera-Guerrero, 2018).

2.3 Data Collection in VR Learning Environment

In order to assess the impact of VR on learning and learning outcomes, data is collected through various methods such as body tracking, performance metrics, physiological responses, questionnaires, and interviews. Body tracking involves analyzing the shape and size of the user's body, and the VR system responds accordingly by adapting the VR environment and the size of objects within it (Olade,Fleming,& Liang, 2020). This may also include VR avatars imitating the body postures and gestures of the users. Performance metric data includes scores from pre-tests, post-tests, and delayed post-tests, as well as data related to task completion time and the number of attempts taken. Physiological sensors are used to collect data on the user's affective state, such as measuring skin conductance levels to evaluate fear and anxiety, heart rate to assess stress, and blood volume pulse amplitude to gauge sympathetic arousal (Feng, González, Amor, Lovreglio,& Cabrera-Guerrero, 2018).
Questionnaires and interviews provide data for quantitative analysis, addressing research questions related to VR usability, user experience, comparisons between VR-aided and non-VR-aided learning, and technology solutions (Hamilton, McKechnie, Edgerton, Wilson, 2021). However, despite the availability of various instruments to collect multimodal data for assessing learning outcomes and system usability, the literature analysis reveals neglect of data related to the learners' interaction behavior.

In non-immersive computer-based learning environments (CBLE), to examine the learners’ learning behavior, the data that is logged has the attributes as mouse-wheel, mouse-wheel click, mouse click left and right, key-stroke, and the mouse movements in addition to the exercise, activity, and timestamp (Rajendran, Munshi, Emara, & Biswas, 2018). Similarly, we developed a mechanism that is able to log all the interaction behavior of the learners happening through HHCs in the VRLE along with timestamps in real-time (Prakash, & Rajendran, 2022; Prakash, Shaikh, & Rajendran, 2023). The designed data collection mechanism is suitable for integration with an immersive VR system where the HMD is connected to a desktop computer. The development of the IBD collection mechanism is discussed in Prakash, & Rajendran, (2022) and Prakash, Shaikh, & Rajendran, (2023). The interaction behaviors are logged in a .csv file in the memory of the computer to which the VR HMD is tethered. The IBD logger contains information such as the HHC used (left or right), buttons used (grip, trigger, control buttons, and thumbstick), button actions (clicked, unclicked, pressed, released, touched, and untouched), button pressure (a value between 0 and 1 indicating the pressure applied), thumbstick axis (x and y co-ordinates), thumbstick angle (a value between 0° and 360°), object interacted and timestamps. An excerpt of the IBD logger is shown in Figure 1. The IBD collection mechanism is deployed in MaroonVR (Pirker, Holly, Lesjak, Kopf, & Gütl, 2019), a VRLE used to learn the concepts of electromagnetic induction. Electromagnetic induction is the phenomenon of inducing electromotive force (emf) by moving a magnet in and around the close proximity of a coil. We used three scenes of Maroon VR: 1) Faraday’s law experiment (the magnet is grabbed and dragged inside the coil to generate emf), 2) the Falling coil experiment (the magnet and an iron bar are allowed to fall freely inside the coil to observe the emf), and 3) the Perspective scene (learners take the perspective of the magnet and generate emf through their walking). The interactions happening when the coil turns (2, 4, and 6 turns), coil diameter (2, 4, and 6 units), and magnetic field strength are varied using virtual interfaces, and the magnet is grabbed, dragged, and dropped and the walking of the learners in the perspective scene are logged in the IBD logger.

Figure 1. An Excerpt of Interaction Behavioral Data Logger

### 2.4 Research Questions

We measured the impact of VR intervention in learning the concept of electromagnetic induction in the VRLE, MaroonVR as shown in Table 1. In this paper, we have used the interaction behavioral data collected to answer the following research questions.

1. Is there a difference in the interaction behavior between the high and low performers?
2. Is it possible to predict learners' learning outcome based on the actions extracted from their interaction behavior?
3. Research Methodology

3.1 Study Design

The VR system utilized in the study (Meta's Oculus Quest 2) includes a precautionary notice indicating that it is not suitable for individuals under the age of 13. Consequently, we refrained from involving school students in our experimentation. Furthermore, since the learning material is already familiar to electrical engineering students, we excluded them from our study. Instead, we opted to conduct our research with a group of fourteen undergraduate engineering students randomly selected from the computer science engineering department, all of whom possess a non-electrical engineering background. After collecting the details related to demography, and familiarity with VR technology from the study participants, the participants’ prior knowledge on the topic of electromagnetic induction was collected using a pre-test. The participants were allowed to play ‘First Steps’, a VR game for approximately 15 minutes to get familiarized with the controllers of the VR system. Then they experienced the MaroonVR VRLE for approximately 30 minutes. The IBD was collected non-intrusively. After the VR intervention, a post-test was conducted to assess the impact of VRLE on the learning outcome. The experiment was conducted with 1 participant at a time and the total study time for a single participant was approximately 1 hour and 15 minutes.

3.2 Analyses

We measured the impact of VR intervention on the learning outcome in our previous publication (Prakash, Shaikh, & Rajendran, 2023). The evaluated result is shown in Table 1.

Table 1. Pre-to-post learning gains - all students (n=14)

<table>
<thead>
<tr>
<th>Pre-test Score Mean (SD)</th>
<th>Post-test Score Mean (SD)</th>
<th>Normalized Gain (SD)</th>
<th>Effect Size (Cohen’s d)</th>
<th>Paired t-test (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.86 (1.75)</td>
<td>7.64 (2.06)</td>
<td>0.42 (0.68)</td>
<td>0.81</td>
<td>3.2 (0.04)</td>
</tr>
</tbody>
</table>

As we intend to explore the impact of interaction behavior in VRLE on the performance, we bifurcated the participants into high and low performers. The participants scoring more than the mean in the post-test are considered high performers and the others as low performers. Accordingly, there are 8 high performers and 6 low performers.

The action events are extracted from the columns of ‘Controller Index’, ‘Button’, ‘Button Action’, and ‘Object’ of the IBD logger to evaluate the difference in the interaction behavior of the high performers and low performers to answer the first research question. The different action events identified from the logged IBD are shown in the Table 2.

Table 2. Action Events extracted from the IBD logger and their description

<table>
<thead>
<tr>
<th>Action Events</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFO</td>
<td>Reading instruction</td>
</tr>
<tr>
<td>NAVIGATE</td>
<td>Teleport, Scene switching</td>
</tr>
<tr>
<td>INTERACT_REL</td>
<td>Handling virtual objects such as magnet and iron bar</td>
</tr>
<tr>
<td>PERS_WALK</td>
<td>Walking in the perspective scene taking the perspective of magnet</td>
</tr>
<tr>
<td>SET_COIL_REL</td>
<td>Setting the turns in the coil between 2 turns, 4 turns and 6 turns</td>
</tr>
<tr>
<td>SET_MAG_REL</td>
<td>Varying the magnetic field strength using VR slider interface</td>
</tr>
</tbody>
</table>

In addition to the interactive actions shown in Table 2, the participants also perform actions which are not logged by the IBD collection mechanism such as walking and turning in falling coil scene and faraday scene, learners’ utterances, and learners’ seeing action. These actions are coined as non-interactive actions.
As mentioned by Prakash, & Rajendran, (2022), the VRLE can be interacted with using specific buttons present in the HHCs such as the Grip button to interact with VR objects and the Trigger button to interact with virtual interfaces. No interactions happen in the VRLE due to the use of irrelevant buttons other than the specified ones and hence they are also kept under the umbrella of non-interactive actions along with the other non-interactive actions such as learners’ utterances and seeing.

A comparative analysis is performed between high performers and low performers to examine the duration and frequency of action events. Several tests are conducted to determine if there are significant differences in prior knowledge, knowledge gained after VR intervention, and the extracted features of action events from the IBD logger. Pearson’s correlation analysis is utilized to identify action events that exhibit a significant relationship with the learning outcome. Following the correlation analysis, multiple linear regression analysis is conducted using the selected action events to answer the second research question. To determine the optimal predictor model, various information criteria, including the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC) scores, are evaluated. This predictor model enables the prediction of performance and the provision of personalized feedback, hints, and learning content to enhance the learning outcome.

4. Results and Discussion

The effect of VR intervention on the learning outcome is evaluated from the data collected using the pre-test and post-test (Prakash, Shaikh, & Rajendran, 2023). We measured the normalized learning gain and effect size to understand the impact of the effect produced by VR intervention. We also found using a paired t-test that the post-test score was significantly higher than the pre-test score with \(t(13)=3.2, p=0.04\), and identified that the significant learning outcome resulted was due to VR intervention (see Table 1).

4.1 Research Question 1: Is there a difference in the interaction behavior between the high and low performers?

Participants are divided into high and low performers to examine how VRLE interaction behavior influences the learning outcome. The homogeneity between the high performers and low performers before VR intervention is assessed using Levene’s test. The results of Levene’s test indicate that there is no significant difference in homogeneity between the high performers and low performers \((p < .05)\) prior to VR intervention \((F = 0.545, p\text{-value} = .474)\). Therefore, the requirement for homogeneity is satisfied.

The significant differences between the high performers and low performers on various parameters were evaluated using a series of Mann-Whitney U-tests. The results are tabulated in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Pre-test Score Mean (SD)</th>
<th>VR Intervention Duration Mean (SD)</th>
<th>Post-test Score Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Performers</td>
<td>5.33 (1.51)</td>
<td>2256.17 (649.17)</td>
<td>5.67 (1.21)</td>
</tr>
<tr>
<td>High Performers</td>
<td>6.25 (1.91)</td>
<td>1743.25 (294.5)</td>
<td>9.13 (0.99)</td>
</tr>
</tbody>
</table>

Mann-Whitney U Test Score

|                | U=16.5, and p-value=.368 | U=13, and p-value=.174 | U=0, and p-value= .002* |

* - significant at p-value < .05

Despite no significant difference in prior knowledge (Mann-Whitney test on pre-test scores), the duration of VR intervention (Mann-Whitney test on VR intervention duration), and the homogeneity of pre-test scores between high and low performers (Levene's test), there is
a significant difference in the learning outcome (Mann-Whitney test on post-test scores). This suggests that the variation in learning outcomes among participants is attributed to differences in their interaction behavior within the VRLE. Consequently, various action events, along with their duration and frequency, were extracted from the IBD logger to investigate the impact of participants' interaction behavior on the learning outcome.

Although there is no significant difference in the duration spent in the VRLE between high performers and low performers, we noticed that high performers exhibited a higher number of relevant interactive actions and spent more time engaging in those actions compared to low performers. Relevant interactive actions are the action events defined in Table 2. We calculated Relevant Interaction Duration % = \( \frac{\text{Total Duration of action events}}{\text{Total Duration of VR Intervention}} \) and Relevant Interaction Frequency % = \( \frac{\text{Sum of frequency of action events}}{\text{Sum of frequency of (action events+non-interactive actions)}} \).

The one-tailed Mann-Whitney U test conducted on the Relevant Interaction Duration % revealed a significant difference at \( p<.05 \) between high performers and low performers, \( U = 9, p = 0.031 \). However, the one-tailed Mann-Whitney U test conducted on the Relevant Interaction Frequency % showed no significant difference at \( p<.05 \) between high performers and low performers, \( U = 24, p = 0.476 \). Thus it is found that there is a significant difference in the duration of relevant interaction between the high and low performers. Whereas, no significant difference is observed in the number of interactions between the high and low performers. The descriptive statistics of the Relevant Interaction Duration % and the Relevant Interaction frequency % for the high performers and low performers are shown in the Figure 2 as a box plot.

![Figure 2. Difference in the interaction behavior between the high and low performers](image)

### 4.2 Research Question 2: Is it possible to predict learners’ learning outcome based on the actions extracted from their interaction behavior?

#### 4.2.1 Correlation Analysis

The difference in the learning outcome between high performers and low performers can be attributed to the interaction behavior of the learners within the VRLE, as indicated by the
results of the Mann-Whitney test and Levene's test. To further explore this relationship, we conducted a Pearson correlation analysis between the action events identified from the IBD logger and the post-test score. The results of the correlation analysis are presented in Table 4.

Table 4. Correlation Analysis Results (Bolded Variables Indicate Significant Correlation with Post-Test Score)

<table>
<thead>
<tr>
<th>Learning Outcome Variable</th>
<th>Action events</th>
<th>Pearson’s r</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-test score</td>
<td>Total VR Intervention Duration</td>
<td>-0.695</td>
<td>.006</td>
</tr>
<tr>
<td></td>
<td>INFO_DUR</td>
<td>0.196</td>
<td>.503</td>
</tr>
<tr>
<td></td>
<td>NAVIGATE_REL_DUR</td>
<td>-0.144</td>
<td>.623</td>
</tr>
<tr>
<td></td>
<td>INTERACT_REL_DUR</td>
<td>-0.068</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>PERS_WALK_DUR</td>
<td>-0.116</td>
<td>.693</td>
</tr>
<tr>
<td></td>
<td>Total Number of Interactions</td>
<td>-0.302</td>
<td>.296</td>
</tr>
<tr>
<td></td>
<td>NAVIGATE_REL_FREQ</td>
<td>-0.626</td>
<td>.017</td>
</tr>
<tr>
<td></td>
<td>SET_COIL_REL_FREQ</td>
<td>-0.487</td>
<td>.078</td>
</tr>
<tr>
<td></td>
<td>NAVIGATE_REL_FREQ</td>
<td>-0.662</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>NAVIGATE_REL_FREQ</td>
<td>-0.344</td>
<td>.228</td>
</tr>
<tr>
<td></td>
<td>INTERACT_REL_FREQ</td>
<td>-0.125</td>
<td>.673</td>
</tr>
<tr>
<td></td>
<td>PERS_WALK_FREQ</td>
<td>-0.116</td>
<td>.693</td>
</tr>
</tbody>
</table>

From the Table 4, the variables Total VR Intervention Duration, NAVIGATE_REL_FREQ, and SET_MAG_REL_FREQ show a significant negative correlation with the post-test score, while the variable Non-Interactive Action Duration exhibits a significant positive correlation. An increase in the values of the negatively correlated variables is associated with a decrease in the post-test score, whereas an increase in the value of the positively correlated variable is linked to an increase in the post-test score.

4.2.2 Regression Analysis

The regression analysis was conducted using the variables having a higher correlation with the post-test score. The forward feature selection algorithm was used to develop multiple linear regression models.

Table 5. Scores of various scales to choose optimum predictor model to predict learning outcome using action events of interaction behavior

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>No. of Predictors</th>
<th>AIC</th>
<th>BIC</th>
<th>HQIC</th>
<th>R2</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Interactive Action Duration, NAVIGATE_REL_FREQ, SET_MAG_REL_FREQ</td>
<td>3</td>
<td>47.669</td>
<td>50.226</td>
<td>-31.906</td>
<td>0.555</td>
<td>1.373</td>
</tr>
<tr>
<td>Total VR Intervention Duration, Non-Interactive Action Duration, NAVIGATE_REL_FREQ, SET_MAG_REL_FREQ</td>
<td>4</td>
<td>49.625</td>
<td>52.821</td>
<td>-29.921</td>
<td>0.360</td>
<td>1.646</td>
</tr>
</tbody>
</table>
To find the best regression model out of the various models developed we evaluated the scores of AIC, BIC, and HQIC (Ventura, M., Saulo, H., Leiva, V., & Monsueto, S., 2019). The model having the minimum score of AIC, BIC, and HQIC is chosen as the optimum model to predict the learning outcome. The AIC and BIC scores are low for the model having the interaction behavior variables such as Non-Interactive Action Duration, NAVIGATE_REL_FREQ, and SET_MAG_REL_FREQ as predictors is considered as Model 1. Whereas, a minimum HQIC score is observed for the model having the variables such as Total VR Intervention Duration, Non-Interactive Action Duration, NAVIGATE_REL_FREQ, and SET_MAG_REL_FREQ as predictors is considered as Model 2. Hence to choose the optimum regression model we evaluated the performance of the two models. The performance of the regression models was evaluated using a training set size of 66% and a test set size of 34%, with 10 iterations of random sampling. Model 1, which used Non-Interactive Action Duration, NAVIGATE_REL_FREQ, and SET_MAG_REL_FREQ as predictors, demonstrated higher R2 (0.555) and lower root mean square error (RMSE) (1.373) compared to Model 2 (R2=0.360, RMSE=1.646). Thus, Model 1 was chosen as the optimum regression model to predict the learning outcome. The scores of various information criterion of the regression models is shown in Table 5.

The result of the multiple regression analysis done on the model 1 chosen as the optimum model is shown in Table 6. The value of $R^2 = 0.55$ for the optimal model indicates that 55% of the variance is explained by the model. The B value in the Table 6 indicates the average change in the post-test score (outcome variable) by 1 unit when the corresponding predictor variables are changed by the given value keeping all the other variables constant. Furthermore, the multicollinearity assumption was tested, and the results indicated that the variance inflation factor (VIF) of all predictor variables was below 10, suggesting that multicollinearity was not violated.

<table>
<thead>
<tr>
<th>Variables</th>
<th>R2</th>
<th>B</th>
<th>p-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Interactive Action</td>
<td>0.555</td>
<td>0.001</td>
<td>0.02</td>
<td>0.244</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAVIGATE_REL_FREQ</td>
<td>0.026</td>
<td>0.044</td>
<td>0.181</td>
<td></td>
</tr>
<tr>
<td>SET_MAG_REL_FREQ</td>
<td>-0.241</td>
<td>0.4</td>
<td>0.182</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

This study examined the interaction behavior of learners in VRLE and identified differences in the duration and frequency of action events between high performers and low performers. The duration of relevant actions performed by participants was found to significantly impact the post-test scores, differentiating high performers from low performers. Additionally, an optimal predictor model was developed using variables related to learners’ interaction behavior, including Non-Interactive Action Duration, NAVIGATE_REL_FREQ, and SET_MAG_REL_FREQ. The predictor model demonstrated performance with an $R^2$ value of 0.555. The study ensured that the variables in the model did not violate the multicollinearity assumption.

Despite the fact that the study was effective in collecting IBD and fitting it using a linear regression model, it was carried out with a lower sample size of 14 patients. Hence, further study with a larger sample size is needed to make the regression model more predictable and transportable. In addition, further work is also required to add the information related to the learner’s views and the pace of interactive actions as important traits in understanding the interaction behavior of the learners in VRLE.

The current research paper has contributed to 1) identifying the key difference between the high and low performers from the perspective of interaction behavior in VRLE, 2) designing a predictor model using significant variables of action events extracted from the interaction behavior, and 3) measuring the impact of VR on learning the subject area of...
electronics engineering as VR studies in electronics engineering are limited. In addition, this study suggests further implications for mining behavioral patterns from the IBD and analyzing the differences in behavioral patterns between high and low performers. Moreover, the study proposes that the interaction behavior comprehended with the performance of the learners can be used by the developers to design VR learner models that can offer adaptive and individualized feedback, hints, and learning content. As a next step, we aim to expand the scope by developing a VR-based Adaptive Tutoring System based on these concepts.

References


Keeping Teams in the Game: Predicting Dropouts in Online Problem-Based Learning Competition

Aditya PANWARa*, Ashwin TSa, Ramkumar RAJENDRANa & Kavi ARYAa

aIDP in Educational Technology, Indian Institute of Technology Bombay, INDIA
adityapanwar@iitb.ac.in

Abstract: Online learning and MOOCs have become increasingly popular in recent years, and the trend will continue, given the technology boom. There is a dire need to observe learners’ behavior in these online courses, similar to what instructors do in a face-to-face classroom. Learners’ strategies and activities become crucial to understanding their behavior. One major challenge in online courses is predicting and preventing dropout behavior. While several studies have tried to perform such analysis, there is still a shortage of studies that employ different data streams to understand and predict the drop rates. Moreover, studies rarely use a fully online team-based collaborative environment as their context. Thus, the current study employs an online longitudinal problem-based learning (PBL) collaborative robotics competition as the testbed. Through methodological triangulation, the study aims to predict dropout behavior via the contributions of Discourse discussion forum activities of participating teams, along with a self-reported Online Learning Strategies Questionnaire (OSLQ). The study also uses Qualitative interviews to enhance the ground truth and results. The OSLQ data is collected from more than 4000 participants. Furthermore, the study seeks to establish the reliability of OSLQ to advance research within online environments. Various Machine Learning algorithms are applied to analyze the data. The findings demonstrate the reliability of OSLQ with our substantial sample size and reveal promising results for predicting the dropout rate in online competition. Overall, the study contributes to online learning by addressing the need to understand and predict dropout behavior in online courses. The study’s methodological triangulation, involving qualitative interviews, provides insights into such contexts’ unique dynamics and challenges by utilizing a fully online team-based collaborative environment.

Keywords: Predictive Models, Machine Learning, Discussion Forum Interaction, Self-Regulated Learning, PBL, Collaboration, Robotics

1. Introduction & Related Literature

There has been a tremendous increase in online courses worldwide for Undergraduate students in engineering and all other fields. COVID-19 has further triggered MOOC enrolment numbers to an exponential level (Baudo & Mezzera, 2021). However, the most prominent drawback of online courses is their high attrition rate; the completion rate usually falls between 2-10% (Maxwell et al., 2018). The success and learning in online courses may be directly attributed to student performance, which can be further attributed to several key factors like content interaction, student satisfaction (Zimmerman, 2012), engagement & activity with course material (Soffer & Cohen, 2019), and motivation (Loizzo, Ertmer, Watson, & Watson, 2017), to list a few. Many such factors can be observed or assessed in a traditional classroom with limited participants. The instructor can guide, adapt and help the ‘failing’ students towards success. All these factors become even more critical when dealing with the online setting where the significant learner interactions are via the resources/content and maybe a discussion or query forum. This necessitates exploring the reasons for dropouts to alleviate this issue and provide an opportunity to inform the instructors to adapt the curriculum/task design, motivate, provide feedback, and devise other ways to keep
success high. It is known that there are numerous benefits of online courses, such as rich content, online asynchronous learning, and low cost. Still, despite that, they continue to suffer from significant dropout rates, hampering pedagogical and economic goals (Quadri & Shukor, 2021). Predicting students' likelihood to complete (or not to complete) a MOOC course, especially from the early weeks, has been one of the hottest research topics in the area of learning analytics (Alamri, Alshehri, Cristea, Pereira, Oliveira, Shi, & Stewart, 2019). Predictive models provide timely information about learners at risk of dropout to inform interventions. Instructors and learners can benefit from the results of the predictive models (Moreno-Marcos, Alario-Hoyos, Muñoz-Merino, & Kloos, 2018), which can help modify and improve course content and pedagogy to reduce dropouts (Moreno-Marcos, Muñoz-Merino, Maldonado-Mahauad, Pérez-Sanaguín, Alario-Hoyos, & Delgado Kloos, 2020). There is a huge potential for data analytics on students' learning processes and outcomes in higher education (Aldowah, Al-Samarraie, & Fauzy, 2019). Learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts to understand and optimize learning and the environments in which it occurs (Siemens & Long, 2011). Students' interaction with online content is crucial and utilized in many research studies. Students' active participation in forums was found to enhance overall performance. The completion level was also closely linked to prior online experience and educational attainment (Meneses & Marlon, 2020). Students' interactions are an important source for understanding the students' behavioral patterns in an online learning process. Other researchers have also claimed that students' social and personal information, such as gender, age, or place, significantly impacted their performance and learning outcomes in general (Mubarak, Cao, & Zhang, 2020). Success in an online course requires a high level of discipline and self-direction, and proper time management to complete the assignments. Thus, certain learning strategies must be developed, and the instructors should encourage the development. Peer Learning should also be encouraged by forming project/study groups wherever possible. Though many prediction models have been developed, they usually don't consider high-level factors like Self-Regulated Learning (SRL), which can greatly impact the learners' success. Prior works have also shown that the lack of SRL skills can be an important factor that leads to failure and dropout in MOOCs (Terras & Ramsay, J. 2015). In a study conducted by a research group at Harvard (Whitehill et al., 2015), multinomial logistic regression was employed to identify students at risk of dropping out of a MOOC. Other algorithms, including Logistic and Linear Regression, Forest Regression, Decision Tree, etc., have also been employed (Liao et al., 2019). It has been noticed that although a decent amount of work has been done in predicting dropouts, they mostly focus on a single stream of data, like learners' interaction with content. The dropout phenomena are studied quantitatively, whereas research emphasizes that dropouts can be highly qualitative and complex (Simpson, 2010). Dropout studies thus lack information on important socio-psychological causes and contingencies (e.g., academic workload, personal experiences, and other commitments). SRL is found to be one of the major factors in online learning. Yet, it is not fully utilized for prediction, so a need is there to use multiple data sources [learner interaction in an online forum, SRL Strategies, Qualitative Studies, learner profile, etc.] as combinations and predictors. There has been quite a lot of research going on in the field of motivation and Learning Strategies, yet there is a research gap in purely online learning scenarios. SRL is usually studied through questionnaires; the Motivated Strategies for Learning Questionnaire (MSLQ) has been widely used. However, it was developed in 1991 and is primarily suited to face-to-face classrooms (Zhou & Wang, 2021). Efforts have been made to modify MSLQ part B for Distance Education (Meijs, Neroni, Gijselaers, Leontjevas, Kirschner, & de Groot, 2019). Still, not many descriptive studies have been conducted to establish the validity. Several other instruments like MSLQ (basic), MSLQ modified, and Self-regulated Online Learning Questionnaire SOL-Q (Jansen, van Leeuwen, Janssen, Kester, & Kalz, 2017), Technology Innovation Questionnaire, NSSE (Pascarella, Seifert, & Blaich, 2010), Educuse Student Technology Survey (Dahlstrom, & Bichsel, 2014), and Self-Regulation for Learning Online (SRL-O) (Broadbent, Panadero, Lodge, & Fuller-Tyszkiewicz, 2022) are also sporadically used in studies. Online Self-Regulated Learning Questionnaire or OSLQ (Barnard, Lan, To, Paton, & Lai, 2009; Bruso & Stefaniak, 2009)
(2016) is one such instrument developed solely for the online environment and accommodates the most important scales from the literature like time management, task strategies, etc. However, many more studies with large sample sizes (Rufini, Fernandes, Bianchini, & Alliprandini, 2021) using OSLQ are warranted.

Several dropout prediction models are created, but they are carried out as posthoc analysis, which fails to ‘anticipate’ the dropouts; early predictions and sequentially suitable intervention strategies are thus the need of the hour. Moreover, the research almost always involves individual participants, so there is a huge shortage of team or collaborative research. The ability to predict a team's performance can have a strong pedagogical potential that has not yet been explored sufficiently in the related literature (Giannakas, Troussas, Voyiatzis, & Sgouropoulou, 2021). Furthermore, when dealing with questionnaires, issues related to self-reported data are always present; this can be somewhat circumvented by conducting longitudinal studies and qualitative interviews. Through this study, we are thus trying to tackle these gaps via

1. Collecting self-reported self-regulated learning (SRL) data through a questionnaire and learner activity data from the discussion forum,
2. Analyzing the SRL data to understand its implications within a team-based problem-based learning (PBL) environment,
3. Developing Early Prediction models specifically designed for an online team-based setting to reduce attrition,
4. Augmenting the investigation with qualitative interviews conducted with teams/team members.

Based on these aims, we propose the following Research Questions:

**RQ1**: How reliably can the Online Strategies for Learning Questionnaire (OSLQ) demonstrate the level of Self-Regulated Learning (SRL) in the context of a fully online problem-based learning (PBL) competition?

**RQ2**: How effective are the early prediction models based on discussion forum interactions and the OSLQ in anticipating and preventing team dropouts in the online PBL environment?

In the upcoming sections of this study, we will describe the study environment and discuss the data collection methods, specifically the utilization of the OSLQ (SRL Questionnaire) and the incorporation of Discourse discussion forum activity data like Topics visited, Posts read, likes received, and given. We will then proceed to build our prediction models and present the results. Finally, we will conclude with a comprehensive discussion section that analyzes and interprets the findings, contextualizing them within the broader scope of our research and exploring their implications.

2. **Study Environment**

To address our RQs, we use e-Yantra Robotics Competition (eYRC) as our study environment. One of the authors worked within e-Yantra, which allowed the use of eYRC. e-Yantra is an initiative hosted in IIT Bombay (in CSE Dept.) with the support of the Ministry of Education (MoE) to spread a wide variety of technical skills across the college. e-Yantra’s initiatives complement the existing education system with a “problem-solving” culture through project-based learning (PBL) approach using Simulators & Robotic Kits. eYRC focuses on building not only “HARD” [STEM, Image processing, control systems, ML, embedded systems, Functional programming, etc.] but “SOFT” [teamwork, time-management, dealing with failure, etc.] skills too. eYRC is unique because it is really a MOOC-style competition that trains participants in complex engineering while competing. The students participate as a team of two to four student members from any year and any branch. Participating teams try to solve gamified real-world problem statements called Themes and are abstracted on a static arena. Incepted in 2012 (eYRC-2012), the competition occurs online in two stages, as shown in Figure 1. Stage 1 is usually simulator/programming based, comprising three tasks-Task 0, Task 1, and Task 2, occurring during September-November. Teams are then selected based on their performance and
advance into Stage 2, which involves hardware and learnings from Stage 1. It consists of Task 3, Task 4, Task 5, and Task 6 conducted between December-March. Every team member has access to all resources, the discussion forum, eYRC website (that hosts questionnaires, feedback forms, etc., needed to be filled out individually by each member). The submission of each task is, however, team-based rather than individual. Each theme comprises tasks and constant feedback provided through various online discussion forums, including Discourse, Discord, and Google Hangouts. The selection of eYRC as the research setting is influenced by its unique attributes as a comprehensive online collaborative robotics competition. Furthermore, segmenting problem statements into distinct deadline-driven tasks within eYRC offers an opportune platform for applying algorithms in early prediction endeavors. First, we gather individual members’ SRL questionnaire responses via the website and each member’s interaction data using the Discourse discussion forum. It is crucial to consider that interaction data should be analyzed within the specific time frame assigned to each task, as each task has its own deadline. Semi-structured interviews are also conducted with the eYRC participants for our study to delve deeper into dropout behavior. Then, using methodological triangulation, we try to achieve the study goals.

3. **Research Design & Methodology**

Our study employs methodological triangulation that uses more than one kind of method to study a phenomenon. It is beneficial in confirming findings, more comprehensive data, increasing validity, and enhancing understanding of studied phenomena (Bekhet & Zauszniewski, 2012). This framework underpins our use of the SRL questionnaire, Discussion forum, and qualitative interviews, as detailed in the following sections.

3.1 **Learning Strategies Questionnaire**

For this study, OSLQ: Online Strategies for Learning Questionnaire, which has six sub-scales, namely, **Goal setting, Environment Setting, Task Strategies, Time Management, Help Seeking, and Self Evaluation**, is used. OSLQ is rated via a 5-point Likert Scale, with 1 being *Strongly Disagree* and 5 being *Strongly Agree*. OSLQ is a developed model; hence no Exploratory Factor Analysis (EFA) will be necessary; Confirmatory Factor Analysis (CFA) will be required to establish reliability and address RQ1. CFA enables us to establish construct validity by testing the relationship between observed factors (question items) and latent factors (subscals) proposed by the model. This step is necessary to ensure that subscale scores within the SRL framework can be distinguished and thus reported separately. Jamovi being open-source and cloud-based, was tested for...
performing the CFA. OSLQ was administered to participants via the web portal with barely any modifications to 10011 individual participants representing 2610 teams registered entering Stage 1 of the competition to attempt Task 0 [basic software installation and a supplementary task]. 4907 individual participants, representing 1520 teams, had filled out the questionnaire, and further analysis was performed on this dataset.

3.2 Discourse Discussion Forum Data

To address RQ2, Discourse (an open-source discussion forum) was utilized in the competition. Participants were enrolled here to access various resources, announcements and post their queries. Discourse readily provides forming “groups,” where an instructor can add only certain participants to a certain group relevant to their ‘theme.’ Six separate groups were formed for six Themes in the analyzed year (2022-23). The total number of individual participants from these six themes was 7604 representing 2214 teams.

Discussion forum learner activity was gathered using the Data Explorer feature of Discourse, where full-fledged Structured Query Language (SQL) queries can be written. Queries like Poll Statistics, Lurkers, High Likers, and many more can be explored by simply running them on the cloud. We also leveraged SQL queries to extrapolate each group’s data using their unique group_id. This data was first separately collected for each theme group based on the tasks’ timeline and then later combined for overall analysis. Major features identified from the data were Topics Entered, Posts Count, Likes Given, and Likes Received (see Appendix A). Since these values were individual-based, they had to be further processed to fit the team-based style of the competition. The MAX values of each feature were finalized because of the nature of the competition that even if one team member “visits” the resources, it can be constituted to the whole team. Hence, we finally utilized four features for further analysis and model building for each team: Max Likes received, Max Likes Given, Max Topics Entered, and Max Posts Read. Furthermore, the time interval (deadline) for the above features between tasks was also crucial since we aim for early predictions. For example, for task1 dropout prediction, we only used the features extracted before the task1 deadline, i.e., Task 0 and Task 1. A similar process was followed for other predictions.

3.2.1 Building Machine Learning Models

All algorithms were run in Python using Google Colab. Since the OSLQ data is in Likert Scale, it was first standardized using suitable Python libraries.

3.2.1.1 Procedure

To build our model, we first performed data pre-processing. An overview of the steps:

1. OSLQ data with aggregate (1520 teams) added with the team’s Discourse activity.
2. After this, each team’s task submission data was added. Task submission data either scored (value >=0) or value NULL, representing a team not submitting task
3. NULL value and scores were suitably standardized to 0 (Not Submitted) and 1 (Submitted) using label encoders. Performance or scoring is not considered to keep the model simple (see Appendix B for the Merged dataset description).
4. As our environment resembles a MOOC model, there was a lot of imbalance in the dataset. An example of this is shown in Figure 2; teams selected in Task 0 of the competition.
5. We used SMOTE as the technique to balance the data. SMOTE stands for Synthetic Minority Over-sampling Technique, and it works by generating synthetic examples of the minority class by interpolating between existing examples in the minority class. Similar data balancing is also used to balance the task-wise data since we focus on early prediction. Figure 3 shows Task 0 composition after SMOTE.

1 Appendix Document
6. We then employed competing ML methods: Logistic Regression, Decision Trees, and Random Forest (see Appendix B2 for details).

7. After building the above models based on a training and test set division of 80%/20%, we also evaluated the models for all performance metrics: accuracy, precision, recall, F1-score, and AUC-ROC (see Appendix B2 for details).

In conclusion, the data pre-processing for our machine learning models involved combining OSLQ data with team Discourse activity and incorporating task submission data. We standardized the task submission data to represent submission status. Due to our MOOC-like environment, we encountered dataset imbalance, which we addressed using the SMOTE and employed Logistic Regression, Decision Trees, and Random Forest as competing ML methods. Finally, their performance was evaluated using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

3.3 Qualitative Data

To establish further reliability, a qualitative study was conducted with eYRC participants that were selected from a pool of eYRC 2021-22 (last year’s version) who were not able to perform well in Stage 1 (dropped out before Task 2) but participated again in the current version (eYRC 2022-23) and cracked the Stage 1 to enter the hardware-based Stage 2. Eight participants representing six teams were invited to participate and consented to a study. The overarching objective was to comprehend the team's view of the discussion forum, unravel the causes of early dropout, analyze the role of discussion forum activities, elucidate their contributions, and triangulate this with our Early Prediction models. These interviews were conducted online over Google Meet. Audio and video were both recorded for analysis purposes. To facilitate the discussion with the participants, guiding questions were provided. Each group member was asked these initial questions to further the conversation (see Appendix C). Otter (a speech-to-text app) was employed, and handwritten notes were taken as a backup.

4. Results and Analysis

In this section, we present the results and analysis of the OSLQ (Self-Regulated Learning Questionnaire) data addressing RQ1. Descriptive statistics and Cronbach's Alpha were used to assess the internal consistency of the questionnaire. We also present the CFA results and the adequacy of the model fit to the data as a satisfactory model fit is crucial for the reliability of the questionnaire. We then present the performance and evaluation of our Early Prediction models to address RQ2. Finally, we discuss the analysis of interviews.
4.1 OSLQ Data Analysis

To answer our RQ1, initially, Descriptive Statistics were performed on the resulting data from OSLQ (n = 4907). Internal consistency for all six subscales was determined using Cronbach’s Alpha, as shown in Table 1. All subscales lie in an acceptable range of (0.70 to 0.95) of Cronbach’s Alpha and are sufficiently reliable for analysis.

<table>
<thead>
<tr>
<th>Sub Scales</th>
<th>Cronbach’s Alpha (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal Setting</td>
<td>0.864</td>
</tr>
<tr>
<td>Environment Setting</td>
<td>0.814</td>
</tr>
<tr>
<td>Task Strategies</td>
<td>0.729</td>
</tr>
<tr>
<td>Time Management</td>
<td>0.812</td>
</tr>
<tr>
<td>Help Seeking</td>
<td>0.799</td>
</tr>
<tr>
<td>Self Evaluation</td>
<td>0.846</td>
</tr>
</tbody>
</table>

In performing the CFA, five statistics reflecting fit were reported: the chi-square goodness of fit statistic ($\chi^2$); the ratio of chi-square statistic to degrees of freedom ($\chi^2$/df); the Comparative Fit Index (CFI); the Tucker Lewis Index (TLI) [also known as the Non-Normed Fit Index (NNFI)]; and the root mean square error of approximation (RMSEA). The model fit results are shown in Table 2 and Table 3 below.

<table>
<thead>
<tr>
<th>Sub Scales</th>
<th>Cronbach’s Alpha (α)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal Setting</td>
<td>0.864</td>
</tr>
<tr>
<td>Environment Setting</td>
<td>0.814</td>
</tr>
<tr>
<td>Task Strategies</td>
<td>0.729</td>
</tr>
<tr>
<td>Time Management</td>
<td>0.812</td>
</tr>
<tr>
<td>Help Seeking</td>
<td>0.799</td>
</tr>
<tr>
<td>Self Evaluation</td>
<td>0.846</td>
</tr>
</tbody>
</table>

Table 2. Exact Model Fit

<table>
<thead>
<tr>
<th></th>
<th>χ²</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2924</td>
<td>237</td>
<td>&lt; .001</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Fit Measures

<table>
<thead>
<tr>
<th>RMSEA 90% CI</th>
<th>CFI</th>
<th>TLQ</th>
<th>RMSEA</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>.948</td>
<td>.94</td>
<td>.0531</td>
<td>.0514</td>
<td>.548</td>
<td></td>
</tr>
</tbody>
</table>

Cronbach’s Alpha confirmed the reliability of the questionnaire’s subscales. CFA established the validity of the questionnaire, ensuring that subscale scores within the SRL framework could be reliably distinguished. The fit statistics of the CFA were examined, and a good fit was found. Descriptive statistics, CFA loadings, and Model Fit Indices description are added in Appendix D3.

4.2 Machine Learning Models Performance

From the initial pool of 1520 teams, a dataset comprising 1290 teams was selected for addressing RQ2. Initial model performance was evaluated by constructing models without oversampling using Logistic Regression, Decision Trees, and Random Forest. Subsequently, the model was refined by incorporating minority oversampling for task0, task1, task2, and stage2 as target variables—early Prediction utilized Discussion Forum activities data before each task’s deadline. The results are presented in Table 4, Table 5, and Table 6.

Table 4. Logistic Regression report for task0, task1, task2 and stage2 prediction

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC-ROC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>task0</td>
<td></td>
<td></td>
<td>.49</td>
<td>.95</td>
</tr>
<tr>
<td>0.96</td>
<td>1</td>
<td>.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 Appendix Document
The study employed machine learning models - Logistic Regression, Decision Tree, and Random Forest - to predict task completion and stage advancement in the e-Yantra Robotics Competition. Results indicated varied model performance across tasks and stages. Notably, the Random Forest model consistently outperformed others, demonstrating higher accuracy and AUC-ROC values. For instance, in predicting Task 0 (task0), Random Forest achieved an accuracy of 0.93 and AUC-ROC of 0.65. Similarly, Task 1 (task1) prediction achieved an accuracy of 0.71 and an AUC-ROC of 0.70. In Task 2 (task2) and Stage 2 (stage2) predictions, Random Forest maintained competitive accuracy and AUC-ROC values, solidifying its predictive prowess.

### 4.3 Qualitative Data Analysis

Through the analysis of the interviews, several themes and issues emerged. The participants in the study found the Discourse Discussion Forum highly effective, being more active on it this year, assessing queries before posting. Participants found it beneficial to observe the progress of other participants in the forum. They felt more satisfied and reassured when they discovered that others were facing similar challenges or getting stuck at similar points in the competition while acknowledging helpful solutions with likes and appreciating the forum's structure and navigation. Additionally, they demonstrated improved time management between academics and competition, consistent effort in completing tasks, and reported the benefits of prior participation. Participants also identified issues about the importance of diligently choosing team members and being shy to post their queries in their previous participation, owing to how they’ll be perceived on the forum.
5. Conclusion

This study aimed to predict dropout behavior in an online collaborative PBL competition using a methodological triangulation approach. The study combined data from participating teams’ Discourse discussion forum activities, a self-reported Online Learning Strategies Questionnaire (OSLQ), and qualitative interviews to enhance the results. OSLQ descriptive results revealed that subscales Goal Setting ($\mu=4.06$), Environment Setting ($\mu=4.1525$), Help Seeking ($\mu=4.03$), and Self Evaluation ($\mu=4.06$) were highly evidenced, while Task Strategies ($\mu=3.8225$) and Time Management ($\mu=3.85$) were on the lower side. A reliability analysis of each subscale was performed, and all subscales were found to be within the acceptable range of 0.70 and above, making OSLQ valid and consistent. Using CFA, we established the reliability of OSLQ using a large sample size from various institutions with participants representing various disciplines of Undergraduate Studies in a team-based dynamics. The qualitative interviews also pointed toward participants’ perception of the importance of Time Management as a success indicator. Instructors can thus create some interventions right at the beginning of the competition to encourage participants. For building machine learning early prediction models, we kept the general aim of keeping the models fairly simple and used discussion forum features. The models were built till Stage 1 (or Stage 2 selection) as we focus on Early Predictions. The performance was poor at the start of the competition (task0) but gradually improved. Random Forest and Decision Trees performed better than the Logistic Regression models. Decision Trees and Random Forest seemingly worked well as they incorporated non-linear relationships between the variables. These findings highlight the potential of machine learning models, particularly Random Forest, in predicting task completion and stage advancement in the online Competition. The results contribute to our understanding of the factors influencing the team’s potential early dropouts and can inform the design of interventions to enhance their learning outcomes. The findings of this study have implications for online course design, instructor interventions, and learner support strategies. By identifying teams at risk of dropout early on, instructors can offer timely support, adaptive feedback, and personalized interventions to improve success. However, it is essential to acknowledge some limitations of this study. The self-reported nature of the OSLQ data introduces potential biases and may not capture the full complexity of learners' strategies and behaviors. In the ongoing work, we are exploring the role of individual participants’ data like Year of Study, Prior participation, Engineering Branch, Gender, and other demographic variables that could provide even more insights on team-based research in the online scenario. In conclusion, the contribution of this ongoing study is multifaceted; first, it cements the OSLQ’s validity and reliability through a large sample size which has been missing in the literature. Secondly, it contributes to the growing body of research on predicting dropout behavior in online courses, especially in a fully online longitudinal environment that is also collaborative.

References


https://www.concordia.ca/content/dam/artsci/research/cslp/docs/TIQ2.pdf


CUMTEL
Fostering Students’ Dialogic Engagement with the Use of Visual Learning Analytics as a Teaching Assistant Tool in Primary School Classrooms

Fan CHEN*, Pengjin WANG*, Deliang WANG* & Gaowei CHEN*
*Faculty of Education, The University of Hong Kong, Hong Kong, China
*gwchen@hku.hk

Abstract: Using visual discourse tools can be a valuable approach for teachers to foster academically productive talk in the classroom. However, teachers often pay considerable attention to highly engaged students and fail to create sufficient opportunities for less engaged students to participate in classroom dialogue. This study seeks to enhance learning by incorporating social network analysis to elucidate students’ levels of engagement in classroom discourse. Over a three-week period, this study analyzed both overall and individual student’s dialogic engagement in the classroom. The results demonstrate that as teachers purposefully lead class dialogue and limit their own speech, students gradually speak more. Additionally, those who initially speak less are given more opportunities to engage in classroom dialogue. These results emphasize the significance of using visualization tools to assist teachers in orchestrating and optimizing classroom dialogue.

Keywords: Visual learning analytics, dialogue visualization, dialogic engagement, social network analysis, primary school classroom

1. Introduction

Classrooms are dynamic social and physical environments where teaching and learning occur through interactions between teachers and students. To facilitate effective learning, teachers must organize and manipulate environmental structures, instructional activities, and classroom resources to guide students’ attention and learning, thus allowing students to acquire a wide range of knowledge and skills. However, teachers often encounter considerable challenges when attempting to ensure that students are engaged in learning and discussion in the classroom (Saunders-Stewart et al., 2012). Despite the availability of various support tools in research (e.g., visualization-based scaffolding), few are used in practice, and there is a lack of sufficient strategies to aid teachers in optimizing their teaching approaches (Echeverria et al., 2019; Urhahne et al., 2009). This phenomenon is particularly pronounced in primary school classrooms where students’ learning performance is highly dependent on practices that place high demands on teachers (Wolery & Hemmeter, 2011). More empirical evidence is needed to explore how to support talk-based instruction in primary school classrooms and enable students to engage in learning more effectively (Mavrikis et al., 2019; Urhahne et al., 2009).

Visual learning analytics is a combination of learning analytics and visualization techniques that represent learning and teaching information. This approach goes beyond simply summarizing and visualizing data and presents the classroom learning process in a form that is easy for users to understand (Hsiao & Lin, 2017). These visual approaches can also function as teaching assistant tools. For example, Chen (2020) developed the classroom discourse analyzer (CDA) to aid teachers in reflecting on classroom dialogue and improving classroom teaching practices. They found that visualization-based scaffolding enhanced teachers’ self-efficacy in orchestrating effective classroom dialogue and significantly impacted classroom teaching behaviors (Chen et al., 2020). While this visual
tool had a positive impact on middle and high school classrooms, its effectiveness in elementary schools, where children are more likely to be distracted and disengaged (Downer et al., 2007), requires further investigation. Furthermore, besides focusing on students who are more proficient in the classroom, these analytical methods must also be extended to focus on marginal learners and the participation of the entire class. Therefore, this study attempts to introduce social network analysis to further support classroom teaching and investigate its impact on students’ engagement in classroom dialogue. The research questions are as follows:

RQ1: What are the influences of discourse-based visual tools on the overall dialogic engagement in the classroom?
RQ2: What are the influences of discourse-based visual tools on individual students’ dialogic engagement in the classroom?
RQ3: Do students exhibit significant improvements in their learning performance when discourse-based visual tools are incorporated into classroom learning?

2. Literature Review

2.1 Talk-Based Instruction

Talk-based instruction is beneficial for student learning and skill development, particularly for elementary school students with less learning experience who require more dialogue guidance from teachers to participate in the classroom. According to the “two-thirds rule” uncovered by Flanders (1964), verbal communication dominates classroom practice and its quality is crucial to what and how students learn during lesson time (Khong et al., 2017). Therefore, it is critical for teachers to consciously guide students’ awareness of their behaviors and discourse.

Michaels and O’Connor (2015) put forward a form of talk referred to as academically productive talk (APT), which holds several talk moves that encourage teachers to give students more space to (1) share or clarify their initial ideas (e.g., “Can you say more about it?”), (2) listen carefully to one another (e.g., “Do you repeat his idea in your word?”), (3) deepen their reasoning (e.g., “Why do you think that?”), and (4) think together (e.g., “Do you add on to his idea? Why?”) (Michaels & O’Connor, 2015; O’Connor & Michaels, 2019). APT prompts teachers to encourage students to elaborate, reason, argue, and share thoughts to aid students’ learning. It creates an atmosphere of equal teacher-student dialogue, which helps students to generate new ideas in collaboration, thus deepening their own understanding and communication skills. Studies have shown positive relationships between productive dialogue and students’ academic achievements in mathematics (Chen et al., 2020), English (Vetter et al., 2021), and reading (Barak & Lefstein, 2021).

2.2 Discourse Visualization Tool

Discourse visualization is a technical scaffolding used in the talk-based classroom that employs visual learning analytics to provide a more structured way for teachers to make sense of dialogue information. This approach is widely used in the learning environment to increase students’ collaborative discourse engagement (Celepkolu et al., 2022) and foster diverse students’ science learning (Ryoo & Bedell, 2019). The Classroom Discourse Analyzer (CDA) is this kind of discourse visualization platform (Chen, 2020), as shown in Figure 1. The CDA comprises dynamic, interactive statistical graphs that represent talk moves between teachers and students (Chen, 2020; Chen & Chan, 2022) and social network analysis that models classroom interaction (Mameli et al., 2015; Wolf et al., 2022).
Dynamic interactive statistical graphs use bubbles, tables, and statistics to show classroom discourse information, as shown on the left side of Figure 1. This information includes the number of words, turns, and student-teacher talk patterns (Chen, 2020). Each bubble in the graph represents a participant’s talk, and the size of the bubble corresponds to the number of the participant’s words spoken. The picture in the lower left corner of Figure 1 illustrates turn-taking among participants and represents the number of student-student and teacher-student talks in the real classroom setting.

Social network analysis identifies overall discursive interaction and individual contribution in regular classroom activities, as shown on the right side of Figure 1. In these analyses, ties indicate the quantity of network participants, and relationships show the amount of messages sent and received by each network member (Bokhove, 2016). Network density is a global metric that represents the total number of possible ties and the overall tightness of the network. Generally, the higher the frequency of interaction, the higher the density of the network. The nodal degree involves in-degree ties (the quantity of messages received from others) and out-degree ties (the quantity of messages sent to others) (Mameli et al., 2015). The degree of centrality measures the relevance of a network participant’s amount of dialogues taking place in a classroom (Bokhove, 2016).

2.3 Dialogic Engagement

Learner engagement refers to whether and how students respond to the learning environment, which correlates with academic performance, persistence, and satisfaction (Bergdahl, 2022; Yang et al., 2022). Dialogic engagement is a critical component of classroom engagement that indicates whether learners actively participate in classroom instruction and exchange ideas with peers or teachers. Analyzing discourse could contribute to understanding students’ classroom engagement. Therefore, this study views student-student and student-teacher dialogue as an indicator to reveal classroom engagement.

Social network analysis is widely used to explore classroom interaction. For example, Bokhove (2016) proposed dynamic social network analysis (SNA) to model classroom
interaction through video and transcript data, representing the changes in overall interaction patterns over time. Mameli et al. (2015) utilized a similar approach to identify the discursive patterns of each classroom activity. Previous studies have shown that the single social network presents the overall interaction of the classroom, and adding temporal analysis can reveal the classroom participation of individual learners. Thus, this study attempts to utilize social network analysis to uncover the overall and individual dialogic engagement over time.

3. Methods

3.1 Participants

This study was conducted in a second-grade primary classroom in Hong Kong, China, with the participation of 20 students and their teacher. The teacher holds a master’s degree and is a senior teacher with seven years of teaching experience. Prior to the official launch of the intervention, we informed students’ parents of the interventional plan and obtained their consent. The classroom consisted of 7 boys and 13 girls with an average age of eight years. The teaching content focused on learning basic programming concepts through graphical operations, such as sequences, loops, control, and events.

3.2 Design of the Intervention

The intervention was divided into two stages, as shown in Figure 2. Prior to the intervention, we conducted two workshops to train teachers on the use of APT and CDA. This guaranteed that the teacher could proficiently use APT as the discourse framework during the class and gain basic knowledge of CDA used by teacher-research collaboration before each class. In the follow-up stage, we carried out three sessions of the intervention over the course of three weeks. During each session, we used video cameras to record the complete classroom for approximately 35 minutes per lesson. We then used CDA to analyze and visualize classroom dialogue interaction by converting the audio from the videotape to text. Based on the CDA analysis, we co-designed a plan for instruction improvement, such as providing opportunities for students who spoke less during the previous class or adding some discussion questions to guide students thinking.

3.3 Data Collection and Analysis

This study aimed to investigate the impact of using CDA as a teaching assistant tool to facilitate students’ dialogic engagement. To achieve this goal, video data of the instructor’s teaching were utilized as the primary data source of analysis. The analysis process involved two main steps. First, audio data from the videotape was transcribed into text data with the timestamp, teacher ID, and student ID, using the transcription tool iflyrec. Second, the text data were converted into SNA network data using Python packages such as numpy and networkx, following the methodology outlined in Bokhove (2016) and Mameli et al. (2015) research. In this approach, the sender is considered the source node, the receiver is the target node, and the word number of utterance is the weight of the edge. For example, if
student A spoke to student B for 20 words, this would correspond to the data in SNA as source node A, target node B, and the weight of the edge as 20. Additionally, when the teacher speaks to the whole class, the teacher is regarded as the source node, and all the students in the class are the target nodes. The class responds as a group to the teacher in a similar manner. Based on the SNA data, we examined the dialogic engagement of students as a whole and also analyzed the dialogic engagement of the individual learner.

Six subject-specific tests provided by the teacher were utilized to measure the learning performance of students. We administered a pre-test and a post-test before and after the intervention, respectively, with each test being completed in approximately 10 minutes. To facilitate ease of calculation, we converted test scores to percentages. A paired-sample t-test was employed to determine if any differences existed between the pre- and post-test results.

4. Results

4.1 Overall Dialogic Engagement

Table 1 represents the overall dialogic engagement in the classroom, revealing that the teacher (node 22) was the central node in all three lessons. As the intervention progressed, classroom dialogue interactions gradually increased, as evidenced by the visual graphs and network data, where the network density significantly increased from 4.76% to 30.52%, and the number of arcs grew from 22 to 141. In the first lesson, only 22 talk turns were established in the entire class, while there was a slight increase in the second lesson. In the third lesson, the number of classroom dialogue interactions increased to 141, indicating more opportunities for student-student and student-teacher dialogue. The core nodes correspondingly added student node 7 and node 18.

<table>
<thead>
<tr>
<th>Lesson 1</th>
<th>Lesson 2</th>
<th>Lesson 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Node</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Number of arcs</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>Network density</td>
<td>4.76%</td>
<td>10.82%</td>
</tr>
<tr>
<td>Core nodes</td>
<td>node 22</td>
<td>node 22</td>
</tr>
</tbody>
</table>

In Figure 3, we observe the relationship between the out-degree of node 22 (teacher) and network density. The out-degree of node 22 refers to the number of times the teacher speaks to the students. Interestingly, Figure 3 indicates an inverse relationship between the out-degree of node 22 and network density. Specifically, as the number of times the teacher speaks to the students decreases, the overall network density gradually increases. In the first lesson, the teacher spoke 75 times, and the overall network was relatively sparse with a density of 4.76%; in the second lesson, the teacher spoke 59 times, and the network density
increased to 10.82%; And in the third lesson, the number of teacher explanations decreased to 48, while the network density increased to 30.52%.

![Figure 3. Out Degree of Node 22 (Teacher) and Network Density.](image)

### 4.2 Individual Dialogic Engagement

Figure 4 displays individual students’ dialogic engagement in the classroom during the three lessons. In the first lesson, the average number of times students spoke was 2.5, and the maximum number of times a student spoke was 8 (node 9). Six students did not participate in the classroom dialogue (e.g., nodes 3, 5, 10), and three students participated in only one class talk (e.g., nodes 4, 7, 13). In the second lesson, the average number of times students spoke was slightly lower at 2.35, and the maximum number of times a student spoke was 8 (node 1). Seven students did not speak in class (e.g., nodes 7, 10, 13, 14), and three students spoke only once (e.g., nodes 3, 6, 19). In the third lesson, the average number of times the students spoke was 7.3, and the maximum number of times a student spoke was 20 (node 7). Only one student did not speak (node 18), and three students only spoke once (e.g., nodes 5, 14, 15).

![Figure 4. Individual Dialogic Engagement in the Classroom.](image)

Figure 4 also reveals slight differences in each student’s engagement in classroom dialogue during the three lessons. The participation of some students in classroom dialogue gradually increased (10 students, including nodes 3, 4, 7, 8, 10, 11, 13, 16, 17, 20), while some students’ participation in classroom dialogues decreased (5 students, including nodes 1, 5, 14, 15, 18). Some students’ participation in classroom dialogues fluctuated to a certain extent.
extent (4 students, including nodes 2, 6, 9, 19), while others remained relatively constant (1 student, node 12).

4.3 Learning Performance

Table 2 illustrates the statistical outcomes of overall learning performance, including the pre-test ($M = 62.28$, $SD = 16.08$) and the post-test ($M = 82.35$, $SD = 13.75$). Results from the paired-sample t-test indicated a significant difference ($t = -4.697$, $p < .001$) in learning performance before and after the intervention, with the post-test scores significantly higher than the pre-test scores.

Table 2. Paired-Samples t-test of the Pre- and Post-test Learning Performance

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>20</td>
<td>62.28</td>
<td>16.08</td>
<td>-4.697</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Post-test</td>
<td>20</td>
<td>82.35</td>
<td>13.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5 presents the individual learning performance of students, with more than half of the students demonstrating higher scores on the post-test than the pre-test. When dividing students into low- and high-engagement groups based on the mean value of dialogic engagement during the three sessions, the analysis results revealed that the learning performance of the students in the high-engagement group ($n = 9, M = 84.97$) was slightly higher than that of the students in the low-engagement group ($n = 11, M = 80.21$); however, there was no statistically significant difference in learning performance between the two groups ($t = -.76$, $p = .457$).

Figure 5. Individual Learning Performance of Students.

5. Discussion and Conclusion

The present study aimed to investigate the impact of visual learning analytics as a teacher assistance tool on students’ engagement in classroom discourse. While previous studies have demonstrated that dialogic frameworks or analytics-based tools can benefit classroom instruction (Chen, 2020; Chen & Chan, 2022), there remains a lack of effective scaffoldings to support teachers’ improvement regarding talk-based classroom instruction. In this study, we introduced a richer classroom dialogue analyzer that included interactive statistical and social network graphs that served as a teacher assistance tool to visualize classroom dialogic progress information and provide evidence to support the improvement of classroom instruction. Our findings from social network analyses indicated that visualization-based support contributed to both overall dialogic engagement and individual student engagement. Furthermore, subsequent tests found that students’ academic performance also increased.
From the perspective of overall classroom engagement, we observed that the less the teacher spoke, the more students expressed their thoughts during lessons, as indicated by the growth of network density. The visual graphs may have helped the teacher become more aware of her speech and understand that a high level of teacher talk meant that opportunities for students to speak were more limited. Consequently, she consciously used APT to guide students to think and speak in subsequent classes (Major, Smørdal, et al., 2022; Major, Warwick, et al., 2018). Similar findings can be reported in van der Veen et al. (2017) and Chen et al. (2020), where teachers consciously encouraged and supported student talk, leading to improved communicative performance. The visualization-based tool can further guide teachers and researchers in negotiating how to improve classroom instruction and provide targeted support. To a certain extent, the tool responds to the need for evidence-based support for classroom instruction (Urhahne et al., 2010). As Amarasinghe et al. (2022) have suggested, designing and orchestrating daily classroom scenarios is a demanding task, and advanced technologies, such as learning analytics or data mining, can assist teachers in tackling the complexity involved in their daily work.

Individual student's dialogic engagement improved, with over half of the students demonstrating more engagement in classroom talk during the three lessons, as indicated by the out-degree of students' nodes. However, it should be noted that, given the fixed lesson time, some students may have multiple opportunities to speak, leading to fewer opportunities for others. This highlights the need to consider adding other technical tools to support students' talks when designing activities. For example, online discussion forums could provide students with opportunities to express their ideas (e.g., Ratan et al., 2022); additionally, an adaptive learning system could be incorporated to respond to individual learning needs (e.g., Wu et al., 2017).

Our analysis of learning performance data revealed a statistically significant increase in students' academic achievement from pre- to post-test stages. We also compared student groups with varying levels of dialogic engagement and found that the high-engagement group slightly outperformed the low-engagement group, although this difference was not statistically significant. We will further investigate the relationship between students’ dialogic engagement and their learning performance by considering students’ other characteristics and learning preferences (Bond & Bedenlier, 2019; Halverson & Graham, 2019). In doing so, we can help teachers obtain a better understanding of students’ engagement in classroom dialogue and how this relates to their academic achievement.

6. Limitations and Future Work

This study investigated the influence of teachers (supported by a discourse-based visual tool) on students’ dialogic engagement in the classroom. The findings indicated that this approach has the potential to foster students’ dialogic engagement. However, as this is an ongoing project, there are some limitations to the current study. First, the three-week duration of the intervention made it difficult for us to estimate its long-term impact on student classroom participation. In the coming months, we will continue to conduct extended interventions to explore the impact of visualization tools on both teachers and students. Second, the lack of a comparison group and the small scale of the intervention may limit the generalizability of the research findings. In future research, we will scale up the intervention and explore the impact of discourse-based visualization on both teachers and students. Third, we focused on data sources derived from classroom videos and achievement tests, without considering data sources regarding changes in the participation processes of students or teachers. Consequently, we were unable to conduct a thorough analysis of the impact of visual support on both teachers and students throughout the intervention. In the next stage of the project, we will collect additional sources of data, such as teachers’ reflective logs, classroom observation notes, and interviews with both teachers and students. These additional data sources will provide a more comprehensive understanding of the intervention and improve the accuracy of our overall conclusions.
Acknowledgements

This work was supported by Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221) and by the Innovation and Technology Commission of the Government of the HKSAR (Grant No.: ITB/FBL/7026/20/P).

References


Advancing Education through Stakeholder Engagement: An Evaluation of the Learning Butler Chatbot's Impact on Instructors, and Learners

Chih-Yang PENG\textsuperscript{a}, Su-Hang YANG\textsuperscript{b}, Pei-Yu HO\textsuperscript{a}, Jen-Hang WANG\textsuperscript{c}\textasteriskcentered & Gwo-Dong CHEN\textsuperscript{a}

\textsuperscript{a} Department of Computer Science and Information Engineering, National Central University, Taiwan
\textsuperscript{b} Department of Hospitality Management, Chien Hsin University of Science and Technology, Taiwan
\textsuperscript{c} Research Center for Science and Technology for Learning, National Central University, Taiwan

\textsuperscript{h}harry@cl.ncu.edu.tw

Abstract: This study aims to address the limitation of existing chatbots in education, which are designed only for learners and do not account for the needs of instructors. To solve this problem, the study proposes a learning chatbot called the "Learning Butler," which integrates resources for optimal learning outcomes and is developed based on the situational theory for mobile and classroom learning environments. The effectiveness of the Learning Butler in improving learning outcomes was evaluated through an experiment conducted at a university of science and technology. The results demonstrated that the Learning Butler significantly improved learners' performance compared to the general chatbot, and learners and instructors provided positive feedback on its ability to offer instant suggestions and feedback during dialogues and interactions. This research contributes to the development of effective educational technology by exploring the potential of chatbots in enhancing learning outcomes, with the ultimate goal of improving the quality of education through the application of advanced technologies and innovative pedagogical approaches.

Keywords: Chatbot in education, Social chatbot, Situational learning, Mobile situated learning, Butler

1. Introduction

1.1 Background

Chatbots have become increasingly popular in education due to their ability to integrate teaching content, provide real-time assistance, and enhance learning engagement (Okonkwo & Ade-Ibijola, 2021). However, most existing chatbots for education are designed only for learners and do not account for the needs of other stakeholders, such as instructors, which can limit their effectiveness. In addition, based on the situated learning theory, through dialogue interaction with roles shaped in physical or digital ways, instructors can overcome the limitations of time and space, and provide learners with real situations to apply their classroom knowledge in a low-cost and convenient way (Dede, 2009).

To address this limitation, this study proposes a learning chatbot called the "Learning Butler," modeled after the role of a butler, to serve as a communication bridge among stakeholders. Through the Learning Butler, learners can receive assistance with time management, emotional management, and encouragement to achieve learning goals.

1.2 Research Objectives

In this study, our main objective is to develop a learning dialogue chatbot, i.e., the Learning Butler, designed as a butler role for learning and modeled after the situated learning theory.
We aim to propose design methods for a learning chatbot that can serve as a communication bridge among stakeholders, integrate various resources, and achieve optimal learning outcomes. The ultimate goal is to enhance the quality of education through the application of advanced technologies and innovative pedagogical approaches.

1.3 Research Questions

To achieve the research goals, the research questions are designed to address the following inquiries: (1) Does the learning butler chatbot lead to better improvements in learners’ learning achievements compared to a general chatbot? (2) Does the learning butler chatbot enhance learners’ learning motivations and sense of responsibility to a better extent than a general chatbot?

2. Related Work

2.1 Application of Chatbot in Education

Chatbots have been recognized as a promising tool for improving education by addressing the problem of insufficient manpower (Winkler & Söllner, 2018). Additionally, chatbots have been found to be helpful in language learning, especially for beginners (Yin & Satar, 2020). Pham et al. (2018) designed a chatbot that assisted English language learners with general greetings, answering specific user requests, providing tips for learning content, and sending reminders to users. However, maintaining learners’ enthusiasm for learning has been identified as a challenge in educational chatbot applications (Fryer et al., 2017). Furthermore, research has indicated that the inability to express understanding to users and the lack of emotion in chatbots may hinder their effectiveness in teaching (Gallacher et al., 2018).

2.2 Situational Learning

Situational Learning emphasizes that learners should acquire knowledge and skills in applied contexts, based on the concept of Situated Cognition proposed by Brown et al. (1989). Dede (2009) suggested that virtual environments can be utilized to optimize the implementation of situational learning by simulating learning situations that are not easily accessible in the real world. Cognitive Immersive Language Learning Environment (CILLE) is an example of a virtual environment that combines artificial intelligence with augmented reality technology, allowing learners to learn Chinese by interacting with AI agents in a situational learning environment (Divekar et al., 2021). In contrast, Huang et al.’s (2016) study showed that using mobile devices for situational learning resulted in better learning outcomes. Furthermore, some studies suggest that mobile learning has a lasting impact on learners’ learning outcomes and can induce positive emotions such as excitement and happiness during the learning process for up to six months after the experiment (Demir & Akpinar, 2018).

2.3 Butler and Social Chatbot

Social chatbots have the potential to establish an emotional connection with users and provide a sense of social belonging, which can enhance users’ engagement in longer and deeper conversations with the chatbot (Zhou et al., 2020). While social chatbots have been applied in various fields, their potential in the education field, particularly as companions and emotional support during learning, remains largely unexplored (Liu et al., 2022).

On the other hand, butlers, in general, are known for their excellent communication skills, organizational abilities, discretion, attention to detail, and active listening skills. These qualities make butlers suitable for supervision, companionship, and learning assistance, yet few studies have applied the butler model to design chatbots for the education field. To fill this gap, this study proposes a learning butler chatbot designed to assist learners with time and
emotion management, step-by-step completion of in-class and after-class exercises, and situational learning environments.

3. System Design

3.1 System Structure

The system structure of the learning butler chatbot comprises two platforms: the classroom situational learning system and the mobile situational learning platform. The chatbot provides learners with eight functions through its user interface, including time management, emotion management, positive reinforcement, learning goal prompts, multiple-branch scripts, dubbing management, manual evaluation, and a reward mechanism (Figure 1). These functions are intended to create a positive and supportive learning environment by providing learners with emotional support, time management, and rehearsal opportunities. Learners can interact with the chatbot through the classroom situational learning system, or interact with the practice website and learning butler chatbot via their mobile devices, enabling them to practice and review what they have learned at any time and place.

![Figure 1. System structure.](image)

The general chatbot, in contrast, only has basic dialogue functionality, limiting it to three functions: multiple situational scripts, manual evaluation, and a reward mechanism. For the other five functions, instructors provide support by reminding learners of homework before class, actively understanding learners' emotions, giving verbal encouragement to the whole class, informing learners of learning goals before the class starts, and changing the tone of voice promptly according to learning content or learners' situations. The multiple-branch scripts used in general chatbots only provide situational information through text and audio files, without situational information provided by pictures such as comics, and without the initial evaluation provided by the chatbot.

3.2 Implementation

The implementation of the classroom situational learning system is based on a digital learning theater design (Hu, 2021) and includes both a learning butler chatbot and a general chatbot. To enable effective rehearsal, multiple branch scripts have been added. Additionally, and Microsoft Azure's speech-to-text service is used for preliminary assessment by the learning butler chatbot. Customized features in the voice assistant's database provide reminders, positive reinforcement, emotional support, and feedback to reduce anxiety and enhance learners' confidence during rehearsal.

Regarding the mobile situational learning platform, respectively including the general chatbot and the learning butler chatbot. After selecting the script, learners are briefed on the
learning butler chatbot or general chatbot rehearsal processes. When entering the scenario rehearsal, the learning butler chatbot first informs the learner of the learning goals and the upcoming learning schedule, then provides positive reinforcement, scenario comics, demonstration audio files, and allows learners to record specific lines. Following the scenario rehearsal, the learning butler chatbot offers emotional support to learners (Figure 2).

4. Experiment

4.1 Experimental Subject

In this study, a total of 60 undergraduate students in a Hospitality Japanese course at a local university of science and technology were recruited. The students were randomly assigned to either the experimental or control group, with 30 learners in each group. The experimental group consisted of 16 males and 14 females, while the control group was composed of 10 males and 20 females. Throughout the experiment, the two groups remained completely independent of each other and had no impact on each other's performance. The learning material for the system is a restaurant script compiled in Japanese with the assistance of the instructor of the Hospitality Japanese course.

4.2 Experimental Process

The experimental process (Figure 3) spanned six weeks and consisted of several stages. A pre-test was conducted before the experiment began. Basic knowledge teaching took place in the first week to establish a solid foundation. In the second and third weeks, script teaching...
was conducted, and learners used the mobile situational learning platform. The experimental group used the learning butler chatbot, while the control group used the general chatbot. In the fourth and fifth weeks, learners rehearsed their presentation using the classroom situational learning system. The experimental group used the learning butler chatbot, and the control group used the general chatbot. In the final week, learners presented their achievements using the classroom situational learning system. At the end of the experiment, all learners completed a post-test and questionnaires. Some learners were selected for interviews to gain deeper insights into their experience.

4.3 Research Instrument

To evaluate the effectiveness of the learning butler chatbot, the study employed a pre-test and a post-test before and after the experiment, as well as a questionnaire administered during the post-test. The tests were developed in collaboration with the Hospitality Japanese course instructor and included multiple-choice, true/false, and fill-in-the-blank questions related to the course content, with a maximum score of 100. The pre-test aimed to gauge learners' prior knowledge, while the post-test aimed to measure the impact of different learning chatbots on learning effectiveness, analyzed using ANCOVA. In addition to the tests, a Likert five-point scale questionnaire was used to gather data on learners' perceptions of the learning butler chatbot. The questionnaire included 17 questions covering five dimensions, four of which measured the elements proposed by the ARCS motivational model: attention, relevance, confidence, and satisfaction (Keller, 1987). The fifth dimension assessed learning responsibility and was developed by the researchers after review by experts.

5. Result and Discussion

ANCOVA was employed to compare the experimental and control groups in post-test scores with pre-test as covariate. The results show that the average and adjusted average scores of the experimental group are 82.40 and 81.70, respectively, while those of the control group are 72.16 and 72.85, respectively. Furthermore, there is a significant difference between the two groups ($F = 6.255$, $p = .015 < .05$). The partial eta square value was .099, with a range from .058 to .138, which indicates the effect size is at a medium level (Cohen, 1992). The data indicate that using a learning butler chatbot can significantly improve learners' performance compared to a general chatbot.

The questionnaire, with a Cronbach's Alpha score of 0.986, indicated high reliability. The results of independent sample t-tests revealed no significant differences between the experimental and control groups, implying that the use of a learning butler chatbot did not affect learners' motivation and learning responsibility.

This study conducted follow-up interviews with four learners and the instructor who participated in the experiment. The majority of learners believed that the learning butler chatbot played a supervisory role and increased their interest in learning. However, a small number of learners felt that general chatbots helped them enter a state of focused learning, which may explain why no significant differences were shown in the five dimensions of the post-questionnaire. Regarding the instructor, she emphasized the unique feature of the learning butler chatbot, which is its ability to determine whether learners are using the correct tone and style of Japanese during dialogues and interactions, and offer instant suggestions and feedback.

6. Conclusion

This study introduces a new learning approach using a learning butler chatbot on mobile and classroom situational learning platforms. It integrates resources, enhances communication efficiency, and enables real-time understanding of learners' needs for personalized and adaptive learning. Learners benefit from time and emotion management and a situational...
learning environment, while instructors can conveniently plan, track progress, and provide individualized support. The results of this study show that using a learning butler chatbot significantly improves learning performance without negatively impacting learners' motivation and responsibility. Additionally, the majority of the interviewed learners and instructors provided positive feedback.

For future research, psychological theories and user experience should be considered to meet different learning needs, and more interactions should be integrated, such as adding texts and images to the mobile situational learning platform or incorporating facial recognition into the classroom situational learning system to provide more interactive experiences.

Acknowledgements

This study is supported by the National Science and Technology Council, Taiwan. Grand Number: MOST 111-2410-H-008-012-MY3 and NSTC 112-2811-H-008-006.

References

Hu, L. L. (2021). Design and produce a digital learning theater software architecture that is easy to continue to evolve [Unpublished master's thesis]. National Central University, Taiwan.


Abstract: The integration of ubiquitous technologies with classroom interactions has the potential to foster positive learning habits in students. The instructors at Educational Technology Program in IIT Bombay tackled the challenge of remote learning during the COVID-19 lockdown by conceptualizing the use of the Slack platform to promote learning habits among their graduate students. Grounded in the IDC theory specifically developed for the Asian population, the design of this pedagogical intervention aimed to promote students’ habit of conveying complex ideas in an accessible manner through two learning activities aligned with their interests. Conversation analysis was used to examine instructors’ actions to develop learning habits in students, and member check interviews were conducted with instructors to verify the pedagogical approach and findings. The study demonstrated how the instructors’ pedagogical intervention on the Slack platform supported the development of learning habits among graduate students. The results exhibited the mapping of individual components of the habit loop, i.e., cuing, routine and harmony with the design and outcome of the approach. Students’ satisfaction with routine activities and impatience for creations from the learning activities indicated harmony which eventually leads to habit formation. The continual pursuit of learning activities without instructor involvement also suggested habit activation. The present study contributes to the current understanding of effective pedagogical practices to foster good learning habits by leveraging ubiquitous learning tools.

Keywords: Learning habits, ubiquitous learning tools, Slack platform, IDC theory, habit loop, interest-driven learning activities, graduate students

1. Introduction

Developing good learning habits is crucial for students to become lifelong learners and acquire essential competencies for the 21st century, such as critical thinking, self-regulated learning, problem-solving, and collaborative skills (Chan et al., 2018). Hence, educational institutions must prioritize these competencies to foster students' professional growth by promoting the cultivation of good learning habits.

The Educational Technology (ET) Program at IIT Bombay in India encompasses Masters and PhD students, postdoctoral scholars, academic staff and faculty members. During the COVID-19 lockdown, the instructors in the ET Program at IIT Bombay faced the challenge of enhancing learning without the ability for physical interactions. It was during this time that the instructors conceived the idea of using Slack, a ubiquitous learning tool supported by mobile and computers, to connect the team and promote learning. They recognized the potential of this cloud-based instant messaging platform (Slack Technologies, Inc., n.d.) to foster good learning habits among the PhD and MTech students. Slack is designed for team communication and collaboration, offering real-time channels, direct messages, and calls. It also integrates a range of productivity and collaboration tools (Montrief et al., 2021), making it a versatile platform for remote learning.

The pedagogical design of this intervention was grounded in the interest-driven creator (IDC) theory which was specifically developed in the context of the Asian population (Chan et al., 2018). The Asian education system and curriculum drives students to the habit of rote learning and achieving academic satisfaction with high scores in grade-level exams.
Chan et al. 2018; Kirkpatrick & Zang 2011). Forming interest-driven or skill-based habits of learning to promote professional growth of students have not been sufficiently considered in formal education. Hence, the instructors in this study applied the IDC theory to encourage positive learning habits among their graduate student community utilising the Slack platform. They achieved this by incorporating creative activities that aligned with students' interests into their daily learning routines. The Slack platform was pivotal in cultivating these habits in students among a diverse team of individuals. With the aid of valuable feedback from both instructors and peers, the platform facilitated increased engagement and led to marked improvements in creative idea generation and effective communication of complex information.

This study aims to demonstrate how Slack, as a ubiquitous learning tool, can be used to promote positive learning habits in an academic context. The learning activities were designed to develop the habit of conveying complex information in an accessible manner. Specifically, these activities, entailing the creation and sharing of newsletters and memes, involved fostering critical thinking abilities to break down complex concepts and ideas into simpler, more manageable parts that can be easily understood by different audiences. Hereby, we present a pedagogical approach for promoting learning habits among students on Slack by implementing the IDC theory. This contributes to the existing knowledge of effective educational practices with ubiquitous learning tools.

2. Literature Review

2.1 Interest-Driven Creator Theory

The IDC theory consists of three anchored concepts: interest, creation and habit which are interconnected in various ways (Chan et al., 2018). The theory assumes that learners, when driven by interest, can be engaged in knowledge creation, and the repetition of such processes in their daily learning routines will help in forming interest-driven creation habits for them.

The concept of habit formation is an integral component of the IDC theory. The habit loop suggests that habits are formed through a three-step process: \textit{cuing environment} (arrangement of place, time, people, or incidents), \textit{routine} (repetitive pattern of activities), and \textit{harmony} (Chen et al., 2020). The theory posits that learning activities can be developed as interest-driven creation activities through a rigorous design process that prioritizes the cultivation of students' interests. The creation activities must be sufficient to arouse curiosity, engage students in the activities, and extend the scope of those activities in the pursuit of personal interests. To ensure that these activities have lasting effects, the study proposes that they should be incorporated into the daily routine so that students can develop a habit of learning through daily engagement with these activities.

Several researchers have employed the IDC theory as a guiding framework to design learning activities and curriculum in computer programming education (Kong et al., 2018; Kong & Li, 2016) and mathematics course (Huang et al., 2020). A recent study highlighted the implementation of an IDC experimental school in nurturing teachers' pedagogical practices and encouraging academic growth in students (Looi et al., 2023). An example of an IDC-based learning activity intervention focused on habit development is the Modeled Sustained Silent Reading (MSSR) (Wong et al., 2020). The main objective of MSSR was to assist students in developing a reading habit. The intervention involved exposing students to a "modeled" and "silent" environment repeatedly. Another recent study designed learning activities using the IDC theory for students to develop an interest in podcast English learning, be immersed in the learning process and, by repeating this process in their daily routines, strengthen learning habits (Chen et al., 2022). The findings provided useful insights for podcasters to invest in establishing learners' interest factor and stimulating experiences to improve their continuous learning intention. While the IDC theory has been utilized in various applications that focus on interest and creation loops (Huang et al., 2020; Kong et
2.2 Habit Formation in Academia Using Ubiquitous Learning Tools

Instilling good learning habits in students is highly valued by teachers and parents for their long-term learning and development, as they significantly impact academic achievements and future success (Chan et al., 2018; Ebele & Olofu, 2017). A habit is a pattern of behavior that is repeated regularly and often happens unconsciously. In the context of this study, habit is used to refer to mental processes which become evident as behavioural inclinations when students are continuously interacting with tasks or actions of interest-driven learning (Chen et al., 2020).

Though there is enough research on the role of ubiquitous learning tools, including social media in academia (Aljawarneh, 2020; Dontre, 2021), there is a lack of empirical evidence on the correlation between these learning tools and academic habit formation. Slack is a ubiquitous learning tool predominantly geared towards professional use in businesses and organisations. A recent study investigated the use of Slack as a communication platform to enhance collaborative learning and online activities in online and flexible programming courses for in-service teachers (Rouhani, 2020). The results indicated Slack as one of the factors that possibly contributed to increased activity and learning in the course, specifically through the role of instantaneous feedback and discussions. Though Slack is regularly used by students and educators in higher education for team communication and collaborative learning (Menzies & Zarb, 2020; Ross, 2019), there is a gap in the research that explores the potential use of the platform for fostering habit development. Additionally, none of the previous studies have applied the IDC theory in academic contexts through ubiquitous and collaborative learning tools to develop learning habits.

The instructors in the ET Program at IIT Bombay hypothesized that if students’ curiosity, creativity and routine was stimulated on the Slack platform, they were likely to acquire good learning habits and evolve into enthusiastic creators. In this approach, the instructors aimed to foster habits of conveying complex and creative ideas in an accessible manner among their graduate students. To achieve this goal, utilising the Slack platform, the instructors integrated learning activities that were aligned with students’ pre-existing interests into their daily learning routines. The instructors methodologically harnessed the habit loop components of the IDC theory by finding the right triggers as cues for starting the habit, following a routine and providing repeated tasks with small goals to students, and examining the outcomes of these activities.

3. Pedagogical Design

3.1 Learning Activities

Utilizing the Slack platform as a ubiquitous learning tool, the pedagogical design of this intervention was based on the habit loop components of the IDC theory. As part of this intervention, the instructors created two specific learning activities to develop the habit of conveying complex ideas and concepts in an accessible manner.

**Learning activity 1:** For this learning activity, the instructors tapped on the known interest of their students, i.e., to communicate with each other through social media. In March 2020, an email was received by all PhD and MTech students of the department to engage in the creation and sharing of a daily ET Newsletter on Slack platform. In terms of content creation of these newsletters, the activity was open to include academic and non-academic issues as well as students’ personal experiences. Since this activity was conducted in the period of COVID lockdown, it prompted a multitude of articles with creative and diverse insights from the team members dispersed across the country. On Slack, the
newsletters served as an opportunity to showcase students' ideas, creating a sense of community and collaboration within the team. The articles comprised diverse sections of entertainment including research, fiction, philosophy, cartoons, creativity corner, health, travel, etc. The key goal of this activity was to cultivate the habit of communicating intricate and innovative concepts in a way that is easily understood by a wide cohort. The purpose of fostering this habit was to enable individuals to deconstruct complex ideas or concepts into simpler and understandable terms which can be consumed by a wider audience. This is an extremely important habit in the field of academia which can help individuals become effective communicators and improve their ability to connect with others.

**Learning activity 2:** In the second learning activity, instructors fostered students’ pursuit to create memes as a voluntary activity. Meme creation and sharing was an activity of interest to students which was carried out by them on their personal chat groups. This was noticed by instructors in an informal discussion, after which the students were asked to start sharing these memes on the Slack platform with all team members. Furthermore, an instructor incorporated the idea of EdTech memes activity as a regular practice associated with assessments for his courses, namely *Human Computer Interactions* and *Learning Sciences*. As part of the assessment, students were given specific guidelines for creating memes related to each course element, and they were required to upload their creative outputs on Slack. To create an effective meme, students must extract complex ideas into a brief message and present it in a way that is easily understandable and engaging. This requires not only creativity but also critical thinking skills to convey and determine how best to present it. Through memes creation, the instructors aimed at developing the habit of conveying complex ideas in an accessible and engaging way, which can be a valuable skill in many personal and professional contexts.

### 3.2 Instructor Actions

The intervention focused on nurturing students' interests, fostering opportunities for meaningful creation and supporting the development of good habits on the Slack platform. Hence, the instructors’ actions involved in this approach were categorized into three primary components, namely tapping creative interests of students, creation of learning activities and motivating students to not only initiate but also maintain engagement in learning activities until they become habitual (Fig. 1).

Instructors tailored creation-based learning activities to student interests and context, thereby aiding them in overcoming their initial resistance to building a new habit. Student interests were identified through informal conversations (on Slack or in person) and classroom observations, which was a continuous and implicit effort by the instructors (Fig. 1). Additional design characteristics were employed to facilitate the cognitive engagement of students with the learning activities. These features included creating a focused question to guide the learning activity, developing activities which were relevant, required critical examination, and were open to diverse and valid perspectives.

After posting the learning activities and creating dedicated channels on Slack, the instructors dedicated their efforts towards inspiring and motivating the students to sustain their engagement in the tasks. By sharing detailed written guidelines, instructors ensured that the purpose of the learning activity was clear to the students before the start of the activity which made it goal-directed.

Modelling of learning activities by instructors is crucial to habit development due to the nature of humans (Chan et al., 2018). The instructors modeled their own creations on Slack showing a clear example of how the task and skill should be performed. This helped the students get a better understanding of the expected outcome. Also, when the students witness the instructor getting involved in the learning activity, they are more likely to become motivated to do the same, which enhances their focus and engagement in the task.
Figure 1. The figure illustrates the instructors' pedagogical design categorized into three primary components.

Furthermore, the instructors often demonstrated positive learning behaviour by recognizing the new knowledge gained from other students and providing feedback on their creation. Here, this aspect holds significant importance as these activities foster the development of habits by combining self-practice along with learning from the strengths and weaknesses of others. Instructors also motivated the students through constructive feedback on each creation.

As the process of habit development for complex tasks is an interplay between one’s agency and situational resources (Chen et al., 2020), the instructors ensured that the students were not restricted in their creativity due to technological access or any other constraints. Students were provided with the autonomy to use simplest resources or complex tools to come up with their products. The following excerpt from an instruction provided to students, for learning activity 1, on the Slack platform was an example of such creative freedom: “Good morning all. Thank you all for your contributions in all aspects. For the future broadcasters here is a small reflection: Do it the way you want to and are interested in doing it. The most important part is to have fun and enjoy doing this. It's not necessary to follow any format. Make your own format and try to involve as many people as possible. They may not be able to give you exactly what you want but that is ok they have their own challenges. Take what you get and use your freedom to create anything you like. Most importantly, have fun.”

The instructors chose the Slack platform to facilitate these learning activities and foster a culture of sharing among the students. By encouraging them to share their work with peers, a sense of community and collaboration was created, which served as another motivating factor for the students to perform the tasks diligently and receive feedback.

### 4. Methodology

#### 4.1 Research Question

While the link between habits and learning is widely recognized, there is much less research that investigates how learning habits are formed in various circumstances using ubiquitous learning tools. The aim of this study was to explore the development of learning habits among students on Slack using the IDC theory. Hence, the following research question was...
investigated in this study: *How does the instructors’ pedagogical design, based on the habit loop of the IDC theory, support the development of learning habits on the Slack platform?*

4.2 Participants

The learning activities were offered on the Slack platform to foster learning habits among PhD and MTech students of the ET Program at IIT Bombay. However, these learning activities were also participated by other team members and the alumni community. For the learning activity 1, there were a total of 5 instructors, 30 students, 12 alumni and 10 academic staff members who participated on the Slack platform through creation activities and feedback mechanisms. However, for the learning activity 2, there were 5 instructors, 51 students, 8 alumni and 2 staff members as participants.

4.3 Data Sources and Data Analysis

The data source of this research came from the time-stamped chat logs of Slack, which was created for the team members of the ET Program in 2018. Instructors often initiated casual conversations on different Slack channels to keep the students connected and motivated. In March 2020, the pedagogical approach, based on the IDC theory, was utilised on two new Slack channels named *newsletter_et* and *et_memes*. The chat logs were downloaded from these channels, and organised chronologically along with relevant documents, images and other media shared within the Slack workspace. The number of posts and threads related to the learning activities on Slack were quantified and presented in the results section.

The chats on each channel were screened to filter the threads showing evidence for different components of the habit loop and instructor actions. Conversation analysis (Meredith, 2019) was used to analyse the participation of instructors and students in the conversations and study the structure and specific aspects of these discussions, including instructor and peer involvement, quality of discussions, frequency of feedback, etc. Patterns were examined in the data to find relevant evidence for the research question, involving linguistic resources such as use of word choices, sentence structure and language, emoticons, and team members’ responses. Analysis and patterns were used to map the instructors’ actions aimed at transforming routine activities into learning habits of conveying complex ideas in simple ways. Additionally, member check interviews were conducted with two of the instructors who played significant roles in the design and conduct of these learning activities. Member check was performed to provide any corrections, clarifications, or additional insights on the pedagogical approach and its relation to the findings.

5. Results

The habit loop of the IDC theory, consisting of cuing environment, routine and harmony was used as the framework to design the pedagogy for these activities. The learning activities chosen in the pedagogical intervention were focused to promote the habit of conveying complex ideas and concepts in a way that is easily understood by a wide cohort. Herein, we analyse the interplay of habit loop components of the IDC theory in building this learning habit among the graduate students employing the Slack platform.

5.1 Cuing Environment

In order to encourage students to start with creative activities without hesitation, a cuing environment should be provided (Chen et al., 2020). For both learning activities, there were three primary context cues that were utilised: a) friendly and social environment of Slack, open to team members, which can spark interest in such creations (Fig. 2a), b) the instructors who modeled such interest-driven creation activities and insightful discussions which could be mimicked by the students (Fig. 2b), and c) constructive feedback provided by
instructors and peers which helped students gain perception and motivated them to excel (Fig. 2c).

5.2 Routine

By making interest-driven creation activities a regular part of students’ routine, they are more likely to develop a habit of pursuing their interests (Chan et al., 2018; Wong et al., 2020). The instructors consistently motivated the students to upload their creative outputs from learning activity 1 on the Slack platform by 10 AM each morning. This enabled their peers to read and discuss the same throughout the day. As a result, there were a total of 31 issues of
newsletters that were created over a period of 4 weeks by 15 broadcasters while all 57 participants were involved in active and rich discussions which entailed 221 thread posts. Memes creations were significantly contributed by junior PhD and MTech students (38 students); however, the activity was sporadically participated by all 66 participants, including instructors and alumni. Though there was no specific time allotted for this activity by the instructors, they were continuously encouraged to create and share them on a regular basis. There were a total of 185 memes with 28 thread posts and hundreds of emoticons associated with them from November 2020 to September 2021, when it was the final activity provided by the instructor. The instructors’ encouragement to establish a routine for interest-driven creations and discussions amongst the students led to not only consistent engagement with these activities, but also a high volume of creative output.

5.3 Harmony

In the habit loop, harmony refers to the affective outcome of the routine activity as well as the integration or stabilization of habits (Chen et al., 2020). The participants expressed their satisfaction from their engagement with these learning activities (Fig 3a), which in turn led to pursuing the routine activity whenever there was an opportunity. The participants also stated their impatience in anticipation of the upcoming creations from their peers (Fig. 3b), suggesting that they have developed certain habits.

![Figure 3: Participants expressing a) satisfaction through these learning activities, and b) their anticipation of similar creations, suggesting development of habitual behavior.]

Such habits are likely the results of repeated exposure and active engagement in relevant activities. Both learning activities are still continued to be pursued, sporadically, by the students on Slack platform without indulgence or moderation of the instructor. Over time, the format of the newsletter evolved; however, even after the instructors’ involvement ended, 25 articles, with rich discussions, have been generated on that Slack channel by multiple cohorts over a period of 2 years. Likewise, for memes, 61 additional memes were created and shared on Slack subsequent to the instructor’s last intervention. The sustained production of these creative outputs also indicated clear evidence of habit activation.

6. Discussion

Creating interest and context-driven learning activities for students is crucial in building habits because it provides a meaningful and enjoyable experience for them. While Slack has previously been utilized as a communication tool to enhance collaborative learning and online engagement (Rouhani, 2020), the present study demonstrates an effective use of the platform for developing positive learning habits, by applying the IDC theory, among graduate students. The pedagogical design exhibits how instructor actions can play a crucial role in developing good learning habits using a ubiquitous learning tool. This also contributes towards one of the questions posed by the creators of IDC theory on training teachers to nurture the interests of students, provide opportunities for meaningful creation, and support
the development of good habits (Chan et al., 2018). The pedagogical approach categorised the teacher’s actions into three main components: stimulating the creative interests of students, designing learning activities, and encouraging students to not only start but also sustain engagement in these activities until they become habitual. A crucial element observed during the discussions between instructors and graduate students on their creative outputs was the instructors’ openness to embracing constructive disagreements about varying perspectives. The instructors deliberately embraced this attitude to encourage the development of critical thinking to foster learning good habits in their students.

The study highlighted the role of the Slack platform, instructors’ actions in the design and application of this intervention, and the interplay of the habit loop components in nurturing the habit of conveying complex ideas in a simple and accessible manner among graduate students. The cuing environment was established through a social and friendly environment provided by the Slack platform, where the instructor modelled interest-driven creation activities and insightful discussions. The use of Slack as a platform enabled graduate students to receive constructive feedback from instructors and peers, leading to enhanced engagement and effective communication of complex information. The routine was established through regular practice of sharing creative products from learning activities on the Slack platform. The harmony, or the affective outcome of the routine activity, was evident in the satisfaction of learning, anticipation expressed by the students for the upcoming creations and continual pursuit of learning activities. The findings of this study can provide valuable insights for educators in this digital world, especially in light of the recent review suggesting the crucial role of the schools in developing good habits for students (Ekman et al, 2022).

Recognizing the positive response to IDC practices in Asia (Looi et al., 2023), this work represents a small attempt to build good learning habits in students through its application. Nevertheless, there were a few limitations associated with the study. The study could not control for potential confounding variables that could influence habit formation, such as prior habits, personality traits, and other external factors. Also, the study measured the development of habits over a short period of time of a few months, and the long-term effects of the pedagogy were not assessed. Further research would be needed to explore the sustainability of the approach. Additionally, the study relied on subjective measures such as satisfaction and motivation to assess habit development. To provide more robust evidence, future research would incorporate markers for quantitative analysis to monitor and record the frequency and duration of engagements in the targeted habit-forming activities.

7. Conclusion

The study utilised the Slack platform to demonstrate a pedagogical approach, rooted in the habit loop of the IDC theory, in building students’ learning habits. Cultivating the habit of conveying complex ideas and concepts in a clear and comprehensible manner can greatly benefit graduate students in their career development. This skill not only enhances their proficiency in effective communication, but also improves their ability to establish meaningful connections with others. Leveraging the Slack platform, the instructors supported the development of this learning habit and fostered a culture of peer learning by tapping creative interests of students, designing relevant learning activities and motivating students to regularly engage in these activities. The results of the study have important implications for educators using ubiquitous learning tools to improve their pedagogical practices and cultivate positive learning habits in their students.

Acknowledgements

We would like to thank the Next Education Research Lab for providing the financial support for our project. We are also thankful to the Tata ClassEdge Limited for its support in the presentation of this work. Our gratitude to the faculty members, alumni, staff, and research scholars, from the ET Program at IIT Bombay, who contributed to the study.
References


Using the Self-regulated Based Personalized Online Learning System for Learning Factorization in Mathematics

Thanyaluck INGKAVARA\textsuperscript{a}, Patcharin PANJABUREE\textsuperscript{b} & Wararat WONGKIA\textsuperscript{a}

\textsuperscript{a}Institute for Innovative Learning, Mahidol University, Thailand
\textsuperscript{b}Faculty of Education, Khon Kaen University, Thailand and Digital Education and Learning Engineering Association of Thailand, Thailand
\textsuperscript{*}patchapan@kku.ac.th

Abstract: Developing a self-regulated based personalized online learning system (SPOLS) for learning factorization aimed to provide appropriate learning materials that allow students to learn factorization and control themselves to accomplish their target. The study aimed to investigate the impact of using SPOLS on students’ learning achievement and their perception of online self-regulated ability. The participants consisted of a single group comprising seventy-two eighth-grade students. They were administered pre- and post-tests before and after completing a lesson on factorization. Additionally, they were required to rate their scores on an online self-regulated questionnaire before and after utilizing SPOLS. The results showed that incorporating SPOLS led to a statistically significant increase in the average students’ achievement on the post-test for the numerical factorization compared to their performance on the pre-test. However, using SPOLS for learning polynomial factorization did not elicit a statistically significant change in achievement scores. Moreover, in part of the self-regulated ability, after using SPOLS participants’ scoring on pre- and post-self-regulated questionnaires, Chi-square implies discovering the relationship between two categorical variables. The result showed that seventeen questionnaire items were statistically significant after using SPOLS.

Keywords: Personalized online learning system, self-regulated, factorization, ubiquitous learning, mathematics

1. Introduction

One strength of implementing technology in the classroom was assisting the oversight a general classroom hardly reaches. So, the idea of personalized learning usually appears in the context of online learning systems since it enables access to such a system anywhere, anytime. Personalization’s intent spurred motivational and metacognitive states like positive affect and focused attention (Ingkavara et al., 2022; Panjaburee et al., 2022; Walkington & Bernacki, 2014, 2019). However, the personalized online learning system might not stand alone without a learning approach since the freedom to use of personalized online learning system can cause learning issues. Such as, students delay finishing the course and are lost in peer interaction. It reflects in the lack of ability to control themselves to accomplish the target, says self-regulated learning.

Self-regulated learning (SRL) is the self-process and self-beliefs that enable learners to transform their mental abilities into academic performance skills. With this, SRL is considered the proactive process that one uses to drive themselves to accomplish learning; for instance, goal setting, strategies selection, seeking help, or self-evaluation (Zimmerman, 2008). With all these SRL strategies, it turns a massive amount of work for students to carry and accomplish in the general classroom setting; that is why technology becomes an efficient assistant. So, this raises the challenges of well-combination self-regulated learning strategies into a personalized online learning system to help students achieve in the learning content.
Other points related, as from a decade review articles (2010-2020) of Lai and Hwang reported, there were just five studies from a hundred focused on using the SRL strategy in e-learning for mathematics (Lai & Hwang, 2021). It is well known that Mathematics is a subject that content mainly presents as numerical, symbolic, and variable. Therefore, most students would be difficult even to begin or give up on learning. With this, an understanding of those abstract notions and their concepts is proposed to decrease confused learning and also gain the ability to learn new knowledge related to the previous concept (Bruner & Kenney, 1965). Likewise, quadratic polynomial factorization is one categorized as fundamental in learning Algebra; still, few researchers relied on this content in personalized online learning systems with SRL strategies. Therefore, developing a self-regulated-based personalized online learning system (SPOLS) becomes challenging for this study by referring to an existing design of our previous study (Ingkavara et al., 2022). With two points looking for, does learning achievement differ when students receive the SPOLS for learning factorization in mathematics, and do students' perceptions of their online SRL differ before and after using the SPOLS for learning factorization in mathematics?

2. Development of Self-regulated Based Personalized Online Learning System

2.1 The Self-Regulated Based Personalized Online Learning System (SPOLS)

SPOLS is developed from the principles of learning preference-based learning systems. It states that individuals could receive proper learning materials relevant to their preferences and provide each conceptual learning outcome (Ingkavara et al., 2022). In addition, SRL strategies are added to assist students in achieving their learning, and the using steps are described as follows:

- The system provides learning materials corresponding to personal learning styles.
- All are asked to set personal goals and sequence learning contents (Figure 1).
- Before involving learning activities, a pre-test needs to be done; the system diagnoses learning ability and provides learning status, which allows students to adjust goals along the learning process (Figure 2).
- Complete post-tests after finishing all contents; the system displays getting scores for each (Figure 3).
- Personal learning analysis summarizes the overall learning achievement, outcome comparison between pre-and post-test, success percent of learning achievement, and success percent of learning time given (Figure 4).
- After finishing all lessons, SPOLS allows retesting and relearning with other learning materials.

2.2 E-learning Materials in SPOLS for Learning Factorization in Mathematics

The e-learning materials are typically concerned in two parts. One was about various learning materials, and the second was conceptual knowledge in factorization. Four learning materials were adopted regarding Felder-Silverman’s Model to cover personal learning preferences to fit the online learning system as proposed in a previous study (Panjaburee & Srisawasdi, 2016). There relies on four characteristics likes active-visual, reflective-visual, active-verbal, and reflective-verbal.

For conceptual knowledge, this system provided two main contents (i.e., numerical and quadratic polynomial factorization). Using virtual algebra tiles is the core idea in developing all these learning materials (Fosnaugh & Mitchell, 2014; Garzón & Bautista, 2018). Students acquired factorization knowledge from the relation of rectangle area that displays as a product of two numerical or even specific two-terms (width and length).
3. Research Methodology

The non-experiment with a quantitative approach was applied in this study to investigate students’ achievements before and after the incorporation of SPOLS; at the same time, students’ perception of online SRL ability before and after was also found out. In this study, 72 eighth-grade students used SPOLS in the second semester 2022. Therefore, one could study through SPOLS within a month according to one’s strategy to accomplish all lessons. Firstly, students were asked to take a pre-online self-regulated learning questionnaire (OSLQ) and a pre-learning achievement test (numerical factorization) followed by activities. Then take the post-learning achievement test (numerical factorization) after finishing all lessons. Secondly, the pre-learning achievement test (polynomial factorization) was asked to take before going with polynomial factorization activities. And finished by taking the post-learning achievement test (polynomial factorization) and scoring on the post-OSLQ.

The four multiple-choice scores awarded for each correct answer were used to answer the first research question. There are fifteen test items for numerical factorization content, while twenty-five items for polynomial factorization, say five questions per sub-lessons. In addition, to answering the second research question, The OSLQ was conducted in this study (Barnard et al., 2009). There were six evaluation subscales, twenty-four items with a five-point Likert scale. The internal consistency of the score by subscale values for Cronbach alpha ranged from 0.67 to 0.90, implying a reliable questionnaire. This study conducted back translation to ensure the content and communication validity of the questionnaire.

4. Results

4.1 Learning Achievement

After all, data was collected, descriptive analysis was computed to describe pre-and post-test data sets by generating summaries about data samples. The findings represent two cohorts related to the content study. For the numerical factorization lesson, the results pointed out the increase of mean score in achievement after students use SPOLS for learning (M_{pre}= 10.26, M_{post}= 11.03), and the achievement score showed more relatively consistent in the post-test (S.D._{pre}= 4.624, S.D._{post}= 4.097). While for the polynomial factorization lesson, the mean score
of the post-test a little bit increased (M_{pre} = 14.90, M_{post} = 15.32); in contrast, the relatively consistent show a decreasing in the post-test (S.D.\_{pre} = 8.000, S.D.\_{post} = 8.299).

The Wilcoxon signed-rank test was conducted to see a comparison. In the case of numerical factorization, 33 students had a higher score on the post-test than on the pre-test. However, 25 students had no change in their scores, while 14 had a higher score in the pre-test. A Wilcoxon signed-rank test showed that using SPOLS for learning numerical factorization elicited a statistically significant change in learning achievement (Z = 3.293, p < 0.001). For the polynomial factorization, nearby results indicated that ten students had no change in score. An exciting was that 32 students got higher scores on the post-test while 30 students were on the pre-test. These results are a significant concern since it is about 41% that students perform not as well. A Wilcoxon signed-rank test showed that using SPOLS for learning factorization in polynomials did not elicit a statistically significant change in learning achievement (Z = 0.820, p = 0.412).

4.2 Perception of Online Self-regulated Learning Ability

The students were asked to rate their perception toward OSLQ, pre- and post-using SPOLS for learning factorization. Chi-squares were conducted to discover the relationship between pre- and post-perceptions. Findings revealed that students’ perception was statistically associated with 17 questionnaire items covering five SRL strategies (out of six): goal setting (GS), task strategies (TS), time management (TM), help-seeking (HS), and self-evaluation (SE). The GS strategy resulted in three sub-items that showed statistically significant with $\chi^2 (4) = 9.917, p = 0.042, \chi^2 (4) = 22.445, p < 0.001$, and $\chi^2 (4) = 11.062, p = 0.026$, respectively. Phi value tests of these three sub-items showed the strength of association between the variables at a moderate level ($0.262 \sim 0.395$). Mean students perceived goal setting and maintaining a high academic standard despite an online course, and students agreed more to not compromise on the quality of assignments because this was an online study, respectively. Moreover, the increasing trend of SRL perception after using SPOLS for these three sub-items, with students’ ratings “strongly agree” almost doubled. In contrast, the environment structuring (ES) strategy was the only one that was not statistically significant. Most of the Phi values showed the strength of association between the variables at a weak level. While the surprise was on the TS strategy, students’ perception indicated that after using SPOLS, their perceptions were highest from all sub-items. Four of them showed a statistically significant difference $\chi^2 (4) = 11.371, p = 0.023, \chi^2 (4) = 14.283, p = 0.006, \chi^2 (4) = 16.759, p = 0.002, \chi^2 (4) = 20.350, p < 0.001$, respectively. The most rated frequency in “strongly agree” is that they agreed to practice solving problems on SPOLS to become more proficient.

For the TM strategy, three sub-items showed statistically significant with $\chi^2 (4) = 26.174, p < 0.001, \chi^2 (4) = 10.151, p = 0.038, \chi^2 (4) = 11.585, p = 0.021$. However, there was one sub-item that showed a relatively strong level of association, which most students rated changing from "neutral" to "agree" and "strongly agree" levels after using SPOLS. Mean that students set more time for learning because they know this is a time-consuming activity but might not be strict, like the same time on the same day. In a while, HS strategy, three sub-items showed statistically significant with $\chi^2 (4) = 16.991, p = 0.002, \chi^2 (4) = 26.475, p < 0.001$, and $\chi^2 (4) = 17.427, p = 0.002$, respectively. Thirty-three students rated the "strongly agree" level after using SPOLS. It indicated that they shared problems with their classmates when using SPOLS, so they knew what they were struggling with and how to solve it. Lastly, SE strategy, all sub-items revealed statistically significant, $\chi^2 (4) = 29.478, p < 0.001, \chi^2 (4) = 14.490, p = 0.006, \chi^2 (4) = 10.900, p = 0.028$ and $\chi^2 (4) = 12.559, p = 0.014$, respectively. Considering the rating frequency, the highest frequency, 31, was about students agreeing to summarize online studies to check their understanding of what they learned.

5. Discussion and Conclusion

The pre-and post-factorization test’s descriptive analysis was separated into two contents,
numerical and polynomial factorization. In the numerical factorization, the mean of the post-test is higher than the pre-test ($M_{post} = 11.03$, $M_{pre} = 10.26$), the same as the mean of the polynomial factorization ($M_{post} = 15.32$, $M_{pre} = 14.90$). Consider the mean difference of each content. Numerical factorization content showed 33 students (45.83%) who improved their factorization ability after using SPOLS. In comparison, 14 students (19.44%) got less than before the study, and 25 students (34.72%) still the same these results show statistically significant at $p < 0.001$. In contradiction, the polynomial factorization content was not statistically significant. However, considering the mean difference probably found the truth that most of the students can improve their ability in factoring polynomials by using SPOLS, 32 students (44.44%), but with the nearest of students who showed a decreasing score, 30 students (41.67%) might unclearly to make a strong confirmation that the use of SPOLS can improve the ability in factoring a polynomial. Overall, learning achievement slightly complies with previous studies that using a personalized online learning system can promote learning achievement (Chen et al., 2020; Chu et al., 2021; Panjaburee et al., 2022). However, one point mentioned, referring to the result of the achievement score of polynomial factorization content, is that it would be better further to seek the leading cause of this issue. Further in-depth study, including individual interviews, might concern task strategies and this group's learning path.

Students scored their perception on OSLQ among using SPOLS for learning. Overall, the result pointed to 17 statistically significant items after using SPOLS. However, one SRL strategy was not substantial, ES. Considering the frequency of score rating, it found that most students rated “strongly agree” on the ability to choose a comfortable place for study, but this frequency still decreased from before. So, finding a learning space that lets them occupy their full potential is pretty hard. Truly that an appropriate learning tool for accessing SPOLS is a computer or laptop; this might generate a messy thing for students since it might state in the public area of a house. So, developing a system or activity that fits portable devices, such as smartphones, the tablet can lead to further study. Next, three sub-items of setting goals in learning showed significance, mean to whether short or long-term, they still set goals and not compromising the quality of assignments just because this is an online study. Since SPOLS’s feature allows students to set a goal at the beginning as long as reset to meet their comfortable, it slightly tailors the target, which motivates them to achieve at the end, which corresponds to the point that even this is the online course they will not reduce the quality in learning. This point corresponds to previous studies that technology could support students in setting their learning goals and instantly monitoring their learning process (Lai & Hwang, 2021). For TS strategy's results showed significance in perception generally state conduct notetaking, talking aloud when doing an activity, preparing discussion-question, and practicing more in the system. It is; therefore, SPOLS provides personalized learning material which fits their learning style, which this point might support learning strategies individually. Likewise, other previous studies that provide individual learning material and activity adjusting based on their preference might help students achieve. (Chu et al., 2021; Walkington & Bernacki, 2019). TM strategy, which focuses on setting extra time to learn, showed statistically significant in that SPOLS was not required every fixed time to study each content but was concerned about the ability to control and achieve what was set before. And for the HS strategy relied on how they sought help and exchanged ideas with friends and teachers. This stage reflected that students know themselves in the problem; one feature that helps them in SPOLS is each learning content status, turning awareness to improve until achieved, which is like SE strategy, which mentions the ability to evaluate themselves. It generally describes the perception of SRL toward using SPOLS for learning, as what is known as SRL is a cyclical process in which student plans, monitors their performance, and then reflects on the outcome (Zimmerman, 1986). The cycle can repeat as one uses until achieved, so it does not fit all as linear. That is why it should be tailored for personnel with specific learning paths. Also, technology becomes the answer to this point, enhancing cognitive and expected learning ability.

In summary, using SPOLS for learning factorization in mathematics can enhance learning achievement in numerical content. In contrast, the perception in SRL showed significance in seventeen of the sub-item. Furthermore, two strategies (i.e., task strategies and self-evaluation) showed statistically significant in all sub-item, indicating that using SPOLS
can enhance self-regulated ability, especially in task strategies and self-evaluation.

Acknowledgments

This research project is supported by the National Research Council of Thailand (NRCT): (Contract No. N41A640223).

References


The Effects of Visualization Strategies on Students’ Learning Outcomes in Augmented Reality Contexts

Pei-Shan Tsai*
Teacher Education Center & Graduate Institute of Technological & Vocational Education,
National Taipei University of Technology, Taiwan
* sandra.pstsai@gmail.com

Abstract: This study attempted to investigate the effects of visualization strategies on electrical-machinery course of vocational high school students’ learning outcomes in augmented reality (AR) contexts. Hence, a learning environment combined with visualization strategies and AR technology were developed in this study. To exploring the effects of visualization strategies, a quasi-experimental design was conducted in this study. The participants of this study included 63 seventh-grade senior high school students from two classes in Taiwan. Two classes were randomly assigned into control and experimental groups. Students in control group learned with AR technology while those in experimental group learned with visualization strategies and AR technology. The results revealed that students in experimental group outperformed than those in control group.

Keywords: Augmented reality, visualizations, electrical-machinery, vocational high school

1. Introduction

The emergence of the AR technology has been continuously rebuilding students’ learning experience. One of the major advantages of the AR technology includes its potential to authentically augment people’s experience in the real world (Wu, Hwang, Yang, & Chen, 2018), which may help learners obtain abstract concepts or perceive unobservable phenomena in a multidimensional way (Xu, Su, Hu, & Chen, 2022). Research results have commonly suggested that individuals may benefit from the AR-enhanced learning environment with increased motivation and engagement (Cai, Chiang, & Wang, 2013), positive attitudes (Behzadan & Kamat, 2013), learning gains in language learning (Cai, Pan, & Liu, 2022), and high levels of reading comprehension (Bursali & Yilmaz, 2019). However, while it is believed that students may profit from the AR technology in improving academic achievement (Lu, Liu, Chen, & Hsieh, 2020; Sahin & Yilmaz, 2020), some studies do not reveal such positive effects (Lai & Chang, 2021). Moreover, in a review conducted by Chang, Binali, Liang, Chiou, Cheng, Lee and Tsai (2022), learners seem to show more positive responses (e.g. learning motivation or attitude) in learning arts-related subjects than those in science.

Indeed, most research has shown promising efficacy on the AR technology in promoting students’ learning outcomes (e.g. Chang & Hwang, 2018; Sahin & Yilmaz, 2020), and some research has suggested that the AR technology utilized with the collaborative pedagogical approach may display higher impact on students’ learning outcomes (Garzón Kinshuk, Baldiris, Gutiérrez, & Pavón, 2020). However, it is noted that relative few studies have been found in suggestion of pedagogical strategies in implementing the AR-enhanced educational context. As the importance of visualization strategies have been emphasis in science education (Ainsworth, Prain, & Tytler, 2011; Chang, Quintana, & Krajcik, 2010). Hence, it is worthwhile to investigate the effects of visualization strategies in the AR-
enhanced educational context so that better AR-enhanced learning environment may be established.

2. Methods

2.1 Participants

The participants of this study included 63 seventh-grade vocational high school students from two classes in Taiwan. Two classes were randomly assigned into control and experimental groups. Students in control group (n = 35) learned with AR technology while those in experimental group (n = 28) learned with visualizations and AR technology.

2.2 Research Process

All the students in two groups enrolled in electrical-machinery courses with the same instructor. At first, all the students were asked to fill in pretest to assess their prior knowledge. During the learning activity, the students in experimental group were learned with visualizations and AR technology while those in control group were learned with AR technology. After the learning activity, all the students in two groups were asked to complete a posttest to understand their learning performance.

2.3 Instruments

2.3.1 AR technology

A learning environment combined with visualization strategies and AR technology were developed in this study. The learning materials in text book and AR context were shown in Figure 1 and Figure 2.
Figure 2. The learning materials in textbook.

Figure 3. The learning materials in AR context. Figure 3 shows a student learns with AR technology. The student can not only browse the learning materials in mobile devices, but also interact with the learning materials to acquire more abstract knowledge or to conduct inquiry learning activities.

Figure 4. An example of a student learns in AR context.

2.3.2 Learning Outcomes

The learning outcomes in pre-test and post-test were developed by two vocational high school teachers. Each test consisted of 15 questions by using two-tier tests. The questions were validated by a teacher and a professor for expert validity.

2.3.3 Data Collection and Analysis
To explore the effects of the visualizations on the students’ learning outcomes in AR contexts, this study used the students’ pre-test data as the covariate and post-test data as a dependent variable to run the one-way ANCOVA analysis.

3. Results

A one-way ANCOVA was used to compare the effect of the two teaching strategies while controlling for the pre-test scores. As shown in Table 1, the results revealed that students in experimental group outperformed than those in control group ($F(1,59) = 4.24$, $p < 0.05$). It showed the effects of visualization strategies in AR learning outcomes.

Table 1. ANCOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>275911.20</td>
<td>1</td>
<td>275911.20</td>
<td>233.16*</td>
</tr>
<tr>
<td>Pre-test</td>
<td>122.51</td>
<td>1</td>
<td>122.51</td>
<td>0.45</td>
</tr>
<tr>
<td>Teaching strategies</td>
<td>1152.29</td>
<td>1</td>
<td>1152.29</td>
<td>4.24*</td>
</tr>
</tbody>
</table>

4. Conclusion

This study investigated the effects of visualization strategies on vocational high school students’ learning outcomes in augmented reality (AR) contexts. The results revealed that students in experimental group outperformed than those in control group. It implied the importance of using suitable teaching strategies into innovative technology into instructions. It is suggested that future studies can investigate the effect of combinations of different teaching strategies and AR technology in AR learning contexts.

Acknowledgements

This study is supported by the National Science and Technology Council, Taiwan, under grant numbers MOST 109-2511-H-027-001-MY3.

References


Teaching Analytics with xAPI: 
Learning Activity Visualization with Cross-platform Data

Izumi HORIKOSHI\textsuperscript{a}, Yuko TOYOKAWA\textsuperscript{b}, Kohei NAKMURA\textsuperscript{b}, Changhao LIANG\textsuperscript{b}, 
Rwitajit MAJUMDAR\textsuperscript{a} & Hiroaki OGATA\textsuperscript{a}

\textsuperscript{a} Academic Center for Computing and Media Studies, Kyoto University, Japan 
\textsuperscript{b} Graduate School of Informatics, Kyoto University, Japan 
* horikoshi.izumi.7f@kyoto-u.ac.jp

Abstract: This study explored the possibility of teaching analytics utilizing daily learning log data recorded in xAPI format for class activity visualization. A junior high school reading activity designed to utilize several ICT tools distributed among multiple platforms was visualized. The visualized class activity was shown to a learning designer of the reading activity unit and then asked what she realized when revising the class activity design. As a result, we found that the integrated learning logs processed from the xAPI could visualize the differences in the actual activities compared with the lesson plan and how active the learners were. In addition, when the learning designer saw this visualization, she expressed her desire to change the activity design to reduce the transition of activities during class. Based on these results, we concluded that the daily learning log data recorded in the xAPI could visualize the time and content of the activities performed by the learners. The results show the possibility of capturing and visualizing the progress of a class by cross-platform analysis using xAPI instead of using multiple special sensors.

Keywords: Teaching Analytics, xAPI, Learning Design, Learning Analytics

1. Introduction

Analyzing and visualizing class activities is called teaching analytics and is essential for research and the professional development of teachers (Hoyos & Velásquez, 2020). Several studies have been conducted to capture learning traces from face-to-face learning contexts. The primary approach to this is Multimodal Learning/Teaching Analytics with various types of data, such as audio, video, text, speech, gestures, and eye-gaze (Ochoa et al., 2017; Prieto et al., 2018; Ndukwe & Daniel, 2020; Martínez-Maldonado et al., 2020). However, learning tools have been increasingly used in the daily teaching context which often enable logging of user interactions in the system. Additionally, some learning tools support xAPI, enabling learning log integration and cross-platform analysis (CPA) (Mangaroska et al., 2019). CPA with xAPI is expected to work as an alternative solution for multimodal sensors.

Given this background, this study explored the possibility of utilizing daily learning log data recorded in xAPI for class activity visualization in teaching analytics. We visualized the reading activity, which was designed to utilize several ICT tools distributed across multiple platforms, and examined the insights that the visualization could provide to reflect on and revise the class activity design. The research questions for this study are as follows:

- **RQ:** To what extent can the daily learning log in xAPI format capture class activities for teaching analytics?

2. Methods

2.1 Methods Overview

This study targeted learning activities that had tasks across multiple platforms. Learning activity visualization was performed using the learning log data collected naturally during the learning activity.
Figure 1 presents an overview of this study. LEAF (Learning and Evidence Analytics Framework) is the Integrated Learning Environment in which the learning activities in this study were conducted. The LEAF system consists of a learning management system (LMS; Moodle), e-Book (BookRoll), learning record store (LRS), and learning analytics tools (log palettes) (Ogata, et al., 2018). All learning logs from these platforms are stored in one LRS in the xAPI format, and logs can be recorded using a unified UUID on the LRS.

![Figure 1. Visualization Method for Class Activities with Cross-platform Log Data.](image)

Based on cross-platform integrated log data, we visualized the learning activities of a class. The lesson plan and visualization were shown to a learning designer and a learning analytics researcher who designed the reading activity unit and was asked what they realized for revising the learning design and lesson plan.

### 2.2 Context

This study focuses on a Japanese junior high school English class with the following class design: The target class used a technique called data-enhanced active reading (Toyokawa et al., 2023), which is a reading class that uses e-Books and learning analytics tool to display logs. Figure 2 shows the lesson plan and the learning platforms used on that day.

![Figure 2. Detailed Lesson Plans and Used Learning Tools.](image)

The lesson plan started with routine activities (dictation, 1 min. reading, etc), followed by a pre-quiz. This was followed by pre-reading, which consisted of reading the text and writing guesses and questions in the e-Book memo. Then, as a Pair Work 1 activity, activities were planned to access the dashboard and share guesses and questions with peers by checking the dashboard. In Pre-reading 2, students were expected to read the passage silently and record their reading time with the e-Book timer. They then accessed the active reading dashboard and registered with their words-per-minute (WPM). Next, they
reread the text and highlighted unknown words and essential points with the e-Book marker. Finally, as in Pair Work 2, the dashboard was accessed, and unknown words and essential points were discussed with peers by rechecking the dashboard.

As described above, the lecture utilized a variety of learning platforms and was designed to combine individual and collaborative activities alternately. Further, this class was not designed for the instructor to give a lecture in front of the students, but rather for the students to proceed with the activities themselves according to their teacher’s instructions.

2.3 Interface, verbs, and activities

Figure 3 shows the interfaces of the three platforms used in the learning activities. When users click each button or link, the learning log is stored in the LRS along with the verbs shown on the side. Visualization was performed based on the logs stored in the LRS. At this time, to make the activity granularity suitable for the visualization of this lesson plan, some verbs were adjusted as grouped or divided into several activities as shown in Figure 3.

(a) Learning Management System

(b) e-Book

(c) Learning Analytics Tools

Figure 3. The Interfaces of Learning Platforms and Verbs.
3. Results and Discussion

Figure 4 shows the activities during the 50-min class, aggregated every minute. The size of each dot represents the number of activities. The vertical axis represents the learning activities shown in Table 1, and the horizontal axis represents time. Each cell in the plots represents a student’s activities, and the bottom cell shows the teacher’s activities. Due to space limitations, only 15 learners out of 38 students were randomly selected for Figure 4.

![Figure 4. Learning Activity Visualization from Integrated xAPI Log Data.](image)

3.1 To what extent can xAPI learning log data capture class activities?

Based on these visualizations in Figure 4, we first analyzed to what extent xAPI learning log data can capture and visualize class activities.

(1) How the class proceeded
The plots in the figure indicate who used the tool, when, and how many times. Thus, this visualization can provide an overview of the class’s progress. The results showed the order in which the tools were used, the differences in tool use timing, and the number of times it was used depending on the person. For example, with the integrated xAPI logs capturing and portraying how the class actually proceeded, almost all students in this class first used the e-Book (blue dots), followed by the dashboard (green dots), and the LMS (red dots) at the end of the class. In addition, this visualization captures variations in how many times (size of dots) and how much time (width of a series of dots) each student performed actions.

(2) Deviation from the lesson plan during the actual activity
Comparing the plan and the whole class activities, the LMS activities represented by the red dots were planned at the beginning of the class but were actually completed at the end of the lesson. This can be interpreted as the teacher changing the lesson plan. In addition, it was planned that the students would use the recommendation in the e-Book to guide their access to the dashboard, but no one used it. This could be because the instructions for the recommendation function were not clear, as they did not realize the function. In this way, this visualization can also assist in assessing how well the lesson progressed as planned.

(3) Whether the students could follow the activity and tool transitions
Regarding the activity with the dashboard represented by the green dots, some students accessed it three times as planned, but others accessed it only once, and a few did not
access it even once. One of the reasons for this could be that the dashboard was not a function in the e-Book but was on learning analytics tools, a different platform launched from LMS. Thus, students found it challenging to access it. In addition, the activities using the dashboard were collaborative; therefore, there were hurdles. In other words, this visualization can explain the difficulty of transitions among tools and activities.

(4) Teacher's behavior and deviation from the lesson plan and students’ activity
Finally, even though teachers accessed all the LMS, e-Books, and learning analytics tools, the timing did not necessarily match the lesson plan and that of the students. In addition, the teacher accessed the quiz for LMS (red dots) and the dashboard for learning analytics tools (green dots), and only the reading and timer for the e-Book (blue dots). Thus, we found that even though it is teaching analytics, it is necessary to pay attention to the logs of the teacher and the learner, especially in the case of a class designed in such a way that students develop activities according to the teacher’s instructions rather than being led by the teacher.

3.2 What kind of insight can the visualization give for the improvement of the learning design?

Second, we showed these visualizations to a researcher who designed this reading activity and asked her what she noticed. We also asked where and how she could improve the lesson plan and learning design.

3.2.1. What the learning designer noticed from the visualization

(1) Unexpected number of page transitions and operations
The learning designer pointed out that the number of reading and timer logs was higher than expected. The reading material consisted of only four pages, each of which corresponded to the activities in the lesson plan, as shown in Figure 2. When it was designed, it was supposed to be read only once per page from the front. However, reading logs appeared in a band shape, implying that page transitions are performed multiple times. The designer mentioned two possible interpretations: one is that students worked diligently, and the other is that the material was not easy to follow and required reviews of the previous pages.

(2) Much time spent on individual activity and less collaborative activities
The designer also focused on the difference between e-Book and dashboard operation logs. Considering that the e-Book reflects individual activities and the dashboard reflects collaborative activities in this design, she reflected as follows: "I am happy they worked hard on e-Book activities, but individual activities may have been more difficult than I expected. I would like to make collaborative activities more active, and modify the plan not to focus on individual activities overly."

3.2.2. What suggestions the designer had for improving the lesson plan and learning design

(1) Activity design improvement: transition between activities and platforms
She noted that the designed lecture had activity switching and movement between platforms, especially with some learners not performing the activities as planned; thus, she would like to reduce back and forth between activities and platforms to remedy the problem. As a concrete measure, she would simplify the activities slightly more while being careful not to lose the purpose of each activity.

(2) System design improvement: gather the functions for one activity in one platform
She also mentioned the system improvement. As mentioned in 3.1 (3), the dashboard was not a function in the e-Book, and some students accessed it only once or did not access it even once. Based on this, she said that if the dashboard could be viewed from within the e-Book, the number of students who could view the dashboard would increase, and collaborative activities would be more active.
4. Conclusion and Future Works

In this study, we set the research question as “To what extent can the daily learning log in xAPI format capture class activities for teaching analytics?” To answer the question, we visualized the reading activity, which was designed to utilize several ICT tools across multiple platforms, and examined the insights that the visualization can provide for reflecting on and revising the activity design.

The results showed that the integrated xAPI logs could visualize whether there were any differences in the actual activities compared with the lesson plan and how active the learners were. In addition, the learning designer mentioned her desire to change the activity design to reduce the transition of activities during class. Based on these results, we concluded that the daily learning log data recorded in the xAPI could visualize the learning activities for teaching analytics.

This study had some limitations. First, our method captured and visualized an entire class activity and compared it with a learning design. However, it should be noted that the target class was designed with activities that used tools frequently. For example, if a class uses only e-Books with few tools, or a class that does not use tools at all, the method captures only a portion of the activity. Second, in this study, we presented the visualization results to the learning designer to see how they would interpret them. However, we would like to show this to teachers and observe how they interpret the visualization. In addition, teachers do not necessarily have sufficient data literacy to interpret learning analytics results (Ndukwe & Daniel, 2020). To solve this problem, feedback should have guiding perspectives instead of simply mirroring the activity (Soller et al., 2005).

Although this research is still in an exploratory phase, it shows the possibility of capturing and visualizing the progress of a class using xAPI instead of multiple sensors.

Acknowledgements

This study was supported by NEDO JPNP20006 and JSPS KAKENHI Grant Numbers JP22K20246 and JP23H00505.

References

Effects of Audio and Tactile Biofeedback Based on EEG Attention Levels on University Students’ Relaxation

Shih-Jou YUa, Wai Ki Rebecca CHENGb, Yi-Hsuan CHENa & Jerry Chih-Yuan SUNa*

a Institute of Education, National Yang Ming Chiao Tung University, Taiwan
b University of the Arts London, UK

* jerrysun@nycu.edu.tw

Abstract: This study aims to examine the effects of audio and tactile biofeedback based on EEG attention levels in anti-phishing education on university students’ relaxation. The study developed an attention feedback system to provide learners with audio and tactile biofeedback by collecting learners’ EEG attention signals and converting them to attention levels. The research method employed a quasi-experimental design. The participants were 90 university students who had no prior anti-phishing learning experiences. A random grouping was adopted to divide the participants into a non-immediate feedback group, an audio-immediate feedback group, and a tactile-immediate feedback group. Each participant was required to wear a portable EEG device that was connected to an attentional feedback system to collect their EEG attention and relaxation signals during the learning activity. After the learning activity, participants were asked to complete a post-activity feedback questionnaire. The results showed that the tactile-immediate feedback group displayed a significantly higher level of relaxation as compared to the audio-immediate feedback group. The study suggests that instructors may consider using tactile-immediate biofeedback based on learners’ attention levels to help regulate attention and improve relaxation in online learning environments.

Keywords: Audio and tactile feedback, EEG biofeedback, relaxation, anti-phishing

1. Introduction

Remote online learning became a new norm in education after the COVID-19 pandemic in March 2020. The impact has silently revolutionized the education system to embrace technology-based cloud learning. However, this brings some challenges. According to the trends report from the Anti-Phishing Working Group (APWG), there was a sudden increase in phishing attacks against videoconferencing service provider Zoom (APWG, 2020) where schools worldwide were switching to video conferencing for their remote learning. Another survey from the 2020 Cyber Threats Report (Netwrix, 2020) stated that 33% of educational organizations felt they were at greater cybersecurity risk than pre-pandemic, 89% of them found new security gaps caused by the rapid transition to remote learning, and 50% of educational organizations had experienced phishing attacks. Today, there is growing concern about VPN exploitation and credential stuffing compared to pre-pandemic when malware or phishing was a major concern. Growing demand for cybersecurity awareness training will become an important trend in K-12 schools and higher education institutions.

Previous studies showed that effective anti-phishing education is the key to preventing phishing attacks (Sun et al., 2016). However, many of the anti-phishing materials highly rely on students’ self-learning. The complex digital worlds and other environmental distractions are competing with human’s limited attentional resources and increasing stress. To make self-learning effective, educational psychologists should accelerate the research on innovative...
approaches to assist students' self-regulation of attention and motivation, as well as managing stress in the midst of learning remotely and individually.

A branch of brain-computer interface (BCI) study has emerged that focuses on designing attention-aware systems that assist users in effectively allocating their attentional focus to optimize engagement (Vertegaal et al., 2006). Since the 1970s, BCI has been actively employed as an assistive technology for clinical or therapeutic purposes, and has been applied to “locked-in” patients or children with attention-deficit/hyperactive disorder (ADHD). In recent decades, the development of Electroencephalography (EEG) and the advances in computing power have provided a cost-efficient, safe, and portable approach which not only allows the use of EEG data to understand users’ cognitive states, but also serves as immediate communication with adaptive interfaces of visual, audio, or tactile feedback to influence and augment cognitive functions (Tan & Nijholt, 2010; Vasiljevic & Miranda, 2023). Sun and Yeh (2017) used audio biofeedback based on EEG attention signals to improve learners’ attention, and showed that it was an effective strategy. The EEG biofeedback can help learners quickly achieve deep relaxation (Rydzik et al., 2023). Xu and Zhong (2018) showed that when learners’ attention and relaxation levels are high, this state of mind can help them learn better. Relaxation, also known as meditation (Xu & Zhong, 2018), can also reduce anxiety (Hardt, 2012). Holmes (2019) suggested that understanding how to integrate neuroscience and mindfulness education to train self-regulation and match learners’ learning requirements is an important issue. Therefore, the issue of how to balance attention and relaxation in the learning process needs to be explored. Previous studies indicated that tactile feedback can improve novices’ relaxation in a digital environment, and does not affect attention (Kim et al., 2021) or the performance of visual tasks (Alahakone & Senanayake, 2010). This study aimed to develop an attentional feedback system that provides audio and tactile biofeedback based on EEG attention levels to create a personalized learning environment. The study also examined the effect of audio and tactile biofeedback on university students' relaxation levels. The following research question was addressed in this study and the research model is shown in Figure 1.

Research question: Are there significant differences in relaxation among university students using non-immediate, audio-immediate, and tactile-immediate feedback based on EEG attention levels in online anti-phishing education?

2. Method

2.1 Participants
The study participants were 90 university students from Taiwan, excluding those who had prior learning experience of anti-phishing. The effective sample comprised 90 (100%) students. The participants were randomly divided into three groups: a non-immediate feedback group \( (n = 30) \), an audio-immediate feedback group \( (n = 30) \), and a tactile-immediate feedback group \( (n = 30) \). There were 36 males (40%) and 54 females (60%), with an average age of 22.7 years.

![Figure 1. The research model](image-url)
and a standard deviation of 2.83. In terms of academic areas, 23 participants were from the College of Humanities and Social Sciences (25.56%), 13 were from the College of Electrical Engineering (14.44%), 11 were from the College of Engineering (12.22%), 10 were from the College of Information Science and Technology (11.11%), 10 were from the College of Management (11.11%), nine were from the College of Science (10.00%), and 14 participants were from other colleges (15.56%).

2.2 Experimental procedure

The informed consent process included the researcher explaining the content of the consent form to the participants before the experiment. After the participants understood the content, they voluntarily agreed and signed the consent form. Before starting the experiment, participants needed to put on the portable EEG device for approximately one minute and confirm that it did not make them uncomfortable. In the next step, researchers took the portable EEG paired with the attentional feedback system and tested the biofeedback for approximately two minutes. When the attentional feedback system was completely prepared, participants could start the online anti-phishing learning activity. The learning activity lasted approximately 30 minutes. The EEG attention and relaxation signals were recorded during the learning activity for the three groups. The non-immediate feedback group did not receive any sensory feedback based on their mental attention states. The audio-immediate feedback group received audio biofeedback when their attention level fell below the threshold of focused attention state. The tactile-immediate feedback group received tactile biofeedback when their attention level fell below the threshold of focused attention state. After the learning activity, participants were asked to complete a post-activity feedback questionnaire, which took approximately 10 minutes. The experimental flowchart is shown in Figure 2.

![Figure 2. The experimental flowchart](image-url)
2.3 Instrument

2.3.1 The attentional feedback system

The attentional feedback system was designed and planned by researchers and developed by a technology company. This system was a mobile application based on Android development platforms, and the mobile device used in this study was a Samsung Galaxy S8 smartphone. The attentional feedback system was connected to the portable EEG device by Bluetooth to collect EEG attention and relaxation signals. It provided attention biofeedback based on learners’ attention levels. The attention levels were calculated by the eSense algorithm and the ThinkGear chip from NuroSky (2017). The ThinkGear chip enhanced the raw EEG signals and filtered noise and muscle movement interference. The eSense algorithm converted the raw EEG values into relative attention levels ranging from 0 to 100 to describe the range of brain activity. Besides 0 meaning the signals were unable to be used, other scores were divided into five ranges to determine the attention levels: 1 to 20 indicated strongly lowered levels, 20 to 40 indicated reduced levels, 40 to 60 indicated neutral levels that were similar to the baseline, 60-80 indicated slightly elevated levels, and 80 to 100 indicated elevated levels. In this study, the biofeedback triggered audio or tactile feedback when the attention scores were below 40 and lasted for more than seven seconds. The system controlled the interval between two instances of feedback for at least 30 seconds. This study also used this system to collect relaxation signals and used the same method to estimate the relaxation levels. The user interface and design of audio and tactile biofeedback are shown in Figure 3.

![Figure 3. The design of audio and tactile biofeedback](image)

2.3.2 The post-activity feedback questionnaire

The post-activity feedback questionnaire was modified from Sun and Yeh (2017). The questionnaire includes three issues: use experience, timing and perception of receiving attention feedback, and other perceptions. It aimed to understand learners’ perceptions of using the attentional feedback system, and collected their suggestions to improve the design of the learning activity and the biofeedback system in the future. The non-immediate feedback
group had five items in their questionnaire, whereas both experimental groups, the audio-immediate feedback and the tactile-immediate feedback group, had eight items respectively.

3. Results

This study used the analysis of variance (ANOVA) to analyze the results by the IBM SPSS Statistic 25 software. The sample conformed to the normal distributions based on the criterion of Kline (2011) and did not reject the hypothesis of the equality of variances by Levene’s test \((F(2,87) = 1.07, p = .35)\). According to Table 1, the ANOVA results indicated that the relaxation levels were significantly different among the three groups \((F = 3.70, p < .05, \eta^2 = .08)\). Because the sample sizes of each group were equal, we chose Tukey’s HSD test to analyze the post-hoc comparisons. The results of Tukey’s HSD’s post-hoc comparison show that the relaxation level of the tactile immediate feedback group \((M = 40.72, SD = 16.61)\) was significantly higher than that of the audio immediate feedback group \((M = 30.41, SD = 13.70)\), indicating that learners who used tactile biofeedback based on EEG attention levels could improve their relaxation levels compared to audio biofeedback. Table 1 shows the descriptive statistics for relaxation and Table 2 shows a summary of ANOVA for relaxation for the three feedback groups.

<table>
<thead>
<tr>
<th>Table 1. Descriptive statistics for relaxation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>Non-immediate feedback group ((n = 30))</td>
</tr>
<tr>
<td>Audio-immediate feedback group ((n = 30))</td>
</tr>
<tr>
<td>Tactile-immediate feedback group ((n = 30))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. A summary of ANOVA for relaxation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of variation</td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Between-group</td>
</tr>
<tr>
<td>Within-group</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Note. (A) indicates the non-immediate feedback group; (B) indicates the audio-immediate feedback group; (C) indicates the tactile-immediate feedback group

4. Discussion and conclusion

According to the results, the relaxation levels were significantly different among the three feedback groups, with the tactile-immediate feedback group displaying higher levels of relaxation as compared to the audio-immediate feedback group. Relaxation can reduce anxiety (Hardt, 2012) and stress, and improve learners’ learning process (Holmes, 2019). Therefore, tactile biofeedback was a helpful strategy to remind learners to focus their attention levels without increasing their anxiety or stress, as a higher relaxation level resulted. Additionally, tactile-immediate feedback helps learners regulate their behavior and avoid competing with visual tasks (Alahakone & Senanayake, 2010). The result is also consistent with Kim et al. (2021), who found that tactile feedback can help people adjust their breathing to improve attention, while audio feedback interferes with attention. Participants’ written
feedback aligned with the findings. Examples are: “When the smartphone vibrates, it reminds me to focus on the content on the screen” (Tactile 27); “When the vibration occurs, I will recall my status of attention” (Tactile 10); “The sound playing may startle and cause distraction” (Audio 29); and “The sound is very harsh” (Audio 28). Therefore, this study suggests that instructors can use tactile feedback based on learners’ EEG attention levels to enhance their relaxation levels. Future research can extend to examining learners’ preferences for volume and melody, negative reinforcement issues, and different tactile patterns.

Acknowledgements

This research was supported by the National Science and Technology Council (formerly Ministry of Science and Technology) in Taiwan through Grant numbers MOST 111-2410-H-A49-018-MY4, MOST 110-2511-H-A49-009-MY2, and MOST 107-2628-H-009-004-MY3. We would like to thank NYCU’s ILTM (Interactive Learning Technology and Motivation, see: http://ILTM.lab.nycu.edu.tw) lab members and the students for helping conduct the experiment.

References


Estimating Physical Interactions with Neighboring Student for Detecting Active Learners in the Computer Classroom

Takahiro YOSHINO*, Shin UENO* & Hironori EGII

*aDepartment of Informatics, The University of Electrical-communications, Japan
y2330128@edu.cc.uec.ac.jp

Abstract: In the classroom, some students may isolate themselves, hesitant to seek help due to psychological barriers. These students need active addressing from Teaching Assistants (TAs). Especially in computer classroom due to the types of study. However, there are cases where some classes are arranged with only a few TAs for a large number of students, and it may cause that students who should be approached may be overlooked. In order to support isolated learners, this study estimates the physical interaction with neighboring student based on the learner video in a computer classroom. Analysis of learner video reveals the possibility of using a depth camera to separate movements and estimating interactions. This result can be applied to the TA support system which detects isolated learners.

Keywords: Student interaction, computer classroom, depth camera, learner video

1. Introduction

Teaching Assistants (TAs) are introduced to support classes in universities. One of the roles of TAs in the classroom is to provide support to students. However, TAs do not have much experience about education (Luo, Bellows, & Grady, 2000). They may overlook students who should be assisted. Some students are also hesitant to seek assistance due to psychological barriers (Price et al., 2017). A proactive approach from TAs is needed for these students.

A more proactive approach is taken by implementing the TA supporting system (Imamura et al., 2020). Support from TAs is not necessary when students can resolve issues among themselves. Detecting student isolation would allow for more appropriate support. Many supporting systems in programming classes use learning results and programming code. It is hard to measure student behavior which do not show up in the code. This study considers shooting from above the classroom. Additionally, a depth camera is used to capture the student behavior. As a measure of student isolation, this study estimates physical interactions with neighboring student.

2. Estimating interaction with neighboring students

This study focuses on classes using computer classroom. In the computer classroom, a display is equipped in each student's seat. In order to avoid being behind the display, this study take pictures from above the classroom without any obstacles. Figure 1 shows images that can actually be taken in the computer classroom.
By using a depth camera, the distance from the camera can be obtained as depth data. When shooting from above, changes of depth data show vertical movement. Using this, the system can separate movements that take place directly above the desk, such as keyboarding, from those that take place at eye level, such as pointing at a display or looking into a neighbor’s display.

3. Analytics method

If changes of depth data can be captured, it may be possible to separate movements just above the desk, movements at shoulder height, and movements where the head changes position. In this study, estimating physical interactions is performed using the separation of these movements and the setting of the calculation area.

To capture the movement of students, the value of change in depth data is calculated between each frame per pixel. To separate movements, it is counted the number of pixels in three ranges of the obtained value of depth data changes. In order to capture the movement just above the desk, the movement of the height near the desk, and the movement of the changing head position, each range are defined and the value of change is counted as 30mm to less than 140 mm for "low," 140 mm to less than 400 mm for "middle," and 400 mm or more for "high". The increase in the number of pixels in each range confirms the occurrence of the movement.

As Figure 1 shows, the video data analyzed in this study shows two student side by side. Analyze one student by splitting the area into two parts, left and right. Physical interactions with neighboring student is estimated by some kind of change in the area between students. For example, point to the next display or move body closer to the next seat. These interactions are analyzed by calculating the space between displays.

4. Result

Learner videos ware collected in a programming class at a science and engineering university. Each class were 90 minutes long. Each student can use the computers equipped in the computer classroom. The video captures the actual student behavior. Intel RealSense L515 was used to collect depth video. Data from 15 students were collected in six times of classes in total. Then, analysis was performed on each video data. The analysis results are shown based on the fourth experimental session in which physical interactions with neighboring student were confirmed.

![Figure 2. Calculation results](image_url)
Figure 2 shows the results of the calculation for one learner. The vertical axis shows the percentage of pixels and the horizontal axis shows the time. In the graph, “low” is represented by red, “middle” by yellow, and “high” by blue.

There is a notable area around 20m50s or 22m30s in the graph. The ratio of “high” does not increase relative to the increase of the ratio of “low” in this range. In this moment, the student was reaching for the mouse. Another notable area is around 24m40s or 25m25s in the graph. As the ratio of “low” increases, the ratio of “middle” and “high” increases. In this moment, the student was moving all of body like a postural change. Moreover, the other notable areas are the ranges from around 22m45s to 24m15s and from around 24m35s to 25m20s. The ratio of low is continuously changing in that range. In this time, the student was keyboarding while his face and body turned in the direction of the display.

The calculation results for the other experimental sessions confirmed the same results. At times when the ratio of “high” changed significantly, movements such as postural changes or changing the position of the chair were confirmed. During time when the ratio of “low” was present in a continuously changing, movements such as keyboarding and checking textbooks on the desk were confirmed.

About calculation results in the space between displays, there were four areas where the “high” ratio intermittent changed. Two of these scenes had one student pointing at the other’s display like figure1 (left). As for the other two parts, they were the scene of taking a seat at the beginning of the class and the scene of getting out of their seats at the end of the class. When interactions were not confirmed, the ratio of “high” has not changed. These range changed only the ratio of “middle” or “low”.

5. Conclusion

This study analyzed actual student movement during class by using depth video from above the classroom. The value of change is calculated in depth information between each frame per pixel. Ratios of depth data change were calculated for each calculation area. Ratio changes compared to actual student movement.

Analysis shows that depth data can be used to separate movements. Especially, it has possibility to separate movements just above the desk, movements at shoulder height, and movements where the head changes position. By setting the calculation area appropriately, depth data may be used to estimate physical interactions with neighboring student. However, the data with interactions is few. In addition to that, it cannot be estimated in complex situations. For example, it was observed that students were not analyzed because of going through the calculation area.

In order to estimate the isolated learner, the system has to capture more clearly the movement of individual students. For example, determine the head position and the body orientation by using depth data. We are planning to develop a system to guide TAs to active support for isolated learner.

References

Effect of Active Breaks during e-Learning and Mental Arithmetic Tasks

Masaki KODAIRA*, Tatsuya HAMADA & Hironori EGI
*Department of Informatics, The University of Electro-Communications, Japan
*2330044@edu.cc.uec.ac.jp

Abstract: In this study, we analyzed the effects of taking active breaks during mental arithmetic and e-Learning tasks. Active breaks are defined as breaks that involve light exercise. An experiment was conducted to analyze the effects of active breaks on recovery from fatigue and the effects on the task results. Here, each subject's fatigue was analyzed using the leg movement measurement method. As a result, in the e-Learning task experiment, recovery from fatigue was observed after active breaks. However, improvement in learning effectiveness was not observed. In the mental arithmetic task experiment, recovery from fatigue was also observed after active breaks. In addition, it is possible that the recovery from fatigue enabled the subjects to maintain a high rate of correct responses to the mental arithmetic task. It was suggested that active breaks may be more effective in improving work efficiency when the cognitive load of the task is greater.

Keywords: Active breaks, leg movement measurement, educational support system, fatigue estimation

1. Introduction

In recent years, due to the influence of COVID-19, an increasing number of students are learning at home. In this context, students are expected to take appropriate breaks on their own, rather than at set times as in school. However, it is difficult to assess fatigue objectively to determine whether it is appropriate to take a break. A previous study suggested appropriate break timing based on students' learning performance (Ramachandran, Huang, & Scassellati, 2017), and the results of that study demonstrated that individually determined break times improved test scores compared to students who took breaks at predetermined times. The study included various breaks, such as tic-tac-toe, stretch, and breathing, however, the effects of the content of these breaks were unclear. Thus, we considered that task performance may differ depending on the student's behavior during a break. In this study, we focused on active breaks, which involves performing light exercise, as a break behavior.

Fatigue is a common cause of poor learner performance. And it is important to consider the degree to which fatigue is recovered by breaks. Most fatigue estimation methods utilize facial expressions and eye blinking. These methods employ video cameras and may give pressure to be monitored to the learner. In contrast, leg movement measurements can estimate fatigue without giving pressure to be monitored to the learner because leg movements are measured using a pyroelectric infrared sensor (Hamada, Terui, & Egi, 2022). When using wearable devices, the discomfort caused by the device may affect concentration and fatigue. However, leg movement measurement devices can be used to estimate fatigue without contact; thus, the influence of the device is negligible.

In this study, we analyzed the effects of active breaks on task performance and fatigue recovery. The target tasks included e-Learning and a mental arithmetic task, because such tasks incur sufficient cognitive loads.
2. Methods

In this study, active breaks were performed by walking on a treadmill. Note that the load incurred by active breaks is considered to vary for individuals. Thus, the Karvonen method was used to set the exercise intensity, and active breaks were performed at a treadmill speed that was appropriate for each learner. The Karvonen method can calculate the heart rate to achieve the target exercise intensity based on the learner's resting heart rate and age (Karvonen, 1957).

Here, the learner's fatigue was estimated by measuring leg movements, and this information can be used to estimate the learner's mental condition without burdening on the learner because it can realize contactless acquisition of biometric information. It has been shown that there is a correlation between the learner's leg movements and mental fatigue (Aikawa, Asai, & Egi, 2019).

3. Experiment and Results

Experiment I involved the e-Learning task, and Experiment II involved the mental arithmetic task.

To analyze the effect of active breaks on the learning effectiveness of e-Learning, we focused on the rate of increase in scores on the pretest and post-test. Here, the rate of increase in scores was calculated by dividing the percentage of correct answers in the post-test by the percentage of correct answers in the pretest. The rate of increase in scores was defined as the rate of correct responses on the post-test divided by the rate of correct responses on the pretest. Results-AB (score) represents the results of the test in the range studied after taking active breaks, and Results-PB (score) represents the results of the test after taking passive breaks.

A paired-sample t-test between the active and passive breaks showed no significant difference in the percentage increase in scores ($p = 0.45$, $t = -0.77$, degrees of freedom = 15). However, the mean score of the test corresponding to learning after passive-breaks was greater than that for the test corresponding to learning after taking active breaks. Thus, it is possible that the passive-breaks were more effective in terms of improving the learning effect.

![Figure 1. Experiment flow](image)

In this study, the amount of leg movement was considered to analyze the effect of active breaks on recovery from fatigue on learners during e-Learning. However, leg movements involve many individual differences; thus, our analysis used the value obtained by dividing the amount of leg movements by the average of the individuals. A Wilcoxon signed-rank sum test revealed that the amount of leg movements after active breaks was significantly lower ($p < 0.05$). An example of low leg motion after active breaks is shown in Figure 1 (upper right, leg movements after active-breaks are in red box). Note that this example was taken from a subject in the PB-AB group.
The results of question indicate that many respondents were positive about the fatigue-relieving effects of the active breaks. There were several comments that could be read as “active breaks helped me feel better.”

The score trends for each of the 100 MATH questions for the 10 subjects did not change significantly for any of the subjects. By comparing the results of the mental arithmetic task after breaks and active breaks, we found that all but three subjects showed a slight increase in scores after taking active breaks. These three subjects did not exhibit a significant decrease of greater than 10 points of scores.

It is thought that as the experiment progresses, fatigue accumulated, concentration declined, and the scores decreased. However, most subjects were able to maintain their scores, which suggests that the active breaks had a positive effect on the MATH task.

Similar to the results of experiment I, here, we found that the amount of leg movement after active breaks was low. A Wilcoxon signed-rank sum test revealed that the amount of leg movements after taking active breaks was significantly lower than the amount of leg movements after taking passive breaks ($p < 0.05$). The results of question indicate that many respondents were positive about the fatigue-relieving effects of the active breaks. However, some subjects were negative about taking active breaks and one subject said “I lost my concentration.”

4. Conclusion

In this paper, we analyzed the effects of taking active breaks while performing two types of tasks, i.e., e-Learning and mental arithmetic tasks. The experimental results demonstrated that the active breaks helped recovery from fatigue in both tasks, but active breaks helped improvement task performance in the only mental arithmetic task. Therefore, the difference in the degree of cognitive load between the e-Learning and mental arithmetic tasks suggests that active breaks have a greater improve task performance effect on tasks with greater cognitive load. However, there were several subjects who did not feel the effects of active breaks; thus, it is necessary to determine which type of breaks is most appropriate for each individual learner.

In the future, we would like to develop a system that suggests appropriate break content and timing based on the given task content and level of fatigue.

Acknowledgments

This work has been partly supported by the Grants-in-Aid for Scientific Research (NOs. 21K18496 and 21K02839) by Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

References


Chronotypes of Learning Habits in Weekly Math Learning of Junior High School

Chia-Yu HSU\textsuperscript{a}, Mandukhail OTGONBAATAR\textsuperscript{b}, Izumi HORIKOSHI\textsuperscript{b}, Huiyong LI\textsuperscript{b}, Rwitajit MAJUMDAR\textsuperscript{b} & Hiroaki OGATA\textsuperscript{b}

\textsuperscript{a}Graduate School of Informatics, Kyoto University, Japan
\textsuperscript{b}Academic Center for Computing and Media Studies, Kyoto University, Japan
*hsu.chiayu.25t@st.kyoto-u.ac.jp

Abstract: Learners may have a unique chronotype of learning habits that they have the preferred time of day to work. Even though learner activity extracted from trace data can provide useful and insightful information about their learning habits, there is a lack of tracing habits in daily learning at a school level from learning logs. Therefore, we propose to understand students' chronotypes of learning habits at the K12 level. We investigate the patterns one week ahead of regular tests over the year using learning analytics techniques of clustering analysis. From 92,694 daily logs of the ninth graders in weekly math learning, we find clusters of learning patterns that suggest different chronotypes of learning habits. The findings enable context-aware recommendations for a more authentic learning experience with adaptivity and personalization, which is potential for enhancing pedagogical practices in mobile, contextualized, and ubiquitous learning environments in future research.

Keywords: Learning Habits, Ubiquitous Learning, Trace Data, Learning Analytics, Adaptive Learning

1. Introduction

When it comes to the preferred time of day to work, learners may have a unique chronotype, such as morningness-eveningness preference (Sher, Hatala, & Gašević, 2022). The chronotype of learning habits can be easily assessed by questionnaires with speedy answers. However, the results tend to be arbitrary and subjectively dependent on individual assumptions (Maslenikova, Rotelli, & Monreale, 2022).

As an approach to deal with the above problem, learner activity extracted from trace data can provide useful and insightful information about the amount and type of activity occurring as well as the time and duration of that activity (Ricker, Kozlarski, & Walters, 2020). However, the studies on learning habits are often in a self-paced learning context like a MOOC (Maslenikova et al., 2022; Ricker et al., 2020). There is a lack of understanding of tracing habits in daily learning at a school level from learning logs. Therefore, extracting the chronotypes of learning habits is important to understand students’ learning behaviors at the K12 level.

In this study, we investigate the patterns one week ahead of regular tests over the year using learning analytics techniques of clustering analysis. The objective is to extract learners’ chronotypes of learning habits from their daily logs. We answer the research question: What clusters of learning patterns can be extracted from the daily logs of math learning?

2. Study Context and Collected Data

In this study, we target the learning context of practicing and testing math exercises in a Japanese junior high school. Figure 1 shows the workflow of the learning activities in the study context. On a tablet computer, learners access an e-book reader, BookRoll (Ogata et al., 2015), and practice exercises of a math concept for a week and then take a test of the same concept on Monday in the following week. After the test, learners check the answers and score
the test with their peers by exchanging each other’s tablets. Finally, learners record the score they get and the full score of the test in a Goal-Oriented Active Learning system, GOAL (Li, Majumdar, Chen, & Ogata, 2021). This workflow was implemented repeatedly on a weekly basis from April 2022 to February 2023.

![Figure 1. Workflow of Weekly Math Learning in Study Context of Junior High School.](image)

From the study context, learning logs collected by BookRoll and GOAL accumulate in the Learning Record Store (LRS). We regard 114 ninth graders as the participants of this study. 92,694 learning logs are extracted for the analyses. In terms of the logs, we calculate the indicator of Time Spent for the analyses to answer the research question in this study. The indicator indicates the minutes a learner spends on a PDF file within an hour. To calculate it, we aggregate the operations from the same PDF file and sum up the diff time in minutes on an hourly basis. Therefore, its value ranges from 0 to 60.

3. Analysis and Results

To answer the research question, we find clusters from weekly patterns of practicing before a test. The weekly pattern is indicated by the time spent within the week before a test day. First, we extract the Time Spent with a sum greater than zero since we do not consider the week in which a learner does not spend time practicing an effective pattern before the test. Second, we divide the hours within a day into four timeslots: morning (05:00-11:00), afternoon (12:00-16:00), evening (17:00-23:00), and overnight (00:00-04:00). We sum up the Time Spent in each timeslot. Third, we make a $4 \times 7$ matrix with the standardized values of the time spent ($X_{kl}^{i,j}$) by the learner in the four timeslots from Monday to Sunday of the week before the test day. $X$ is the z-score of the Time Spent, $i$ indicates the timeslot, while $j$ indicates the day of the week. $k$ indicates the matrix of the $k^{th}$ week. Finally, we convert the matrix into a vector with twenty-eight elements.

Regarding the clustering analysis, we input the 723 vectors that imply the weekly patterns. We adopt the Average Silhouette Method to find the optimal number of clusters and identify ten with the greatest silhouette score. Figure 2 shows the ten clusters emerging from the K-means Cluster Analysis. The red line visualizes the center weekly pattern indicated by the vector of the standardized values of the Time Spent in the four timeslots across the week.

In this study, the chronotype is indicated by the maximum value in the weekly pattern. Based on the labels of the ten clusters and their description, we group the clusters of patterns into four chronotypes:

- Morning on Weekdays (n=236): greatest Time Spent in the morning from Monday to Thursday
- Evening on Weekdays (n=285): greatest Time Spent in the evening from Monday to Thursday
- Afternoon on Weekends (n=35): greatest Time Spent on Sunday afternoon
- Mixed (n=167): Time Spent on Thursday afternoon almost equals that on Sunday evening
4. Discussion and Conclusion

Based on the findings, we can give practical suggestions such as the timing to send the notification messages considering the chronotypes of the learning habits. The knowledge of the temporal context of the learning habits can provide valuable insight into the “contextual profiling” mechanism since it will capitalize on the learner’s schedule for engaging in different learning-related activities.

Collectively, we find this study makes contributions by extracting the chronotypes of the learning habits via the clustering method and dealing with innovative ideas about contextual profiling. The former is novel in identifying the chronotypes from a perspective of educational data mining, while the latter enables context-aware recommendations for a more authentic learning experience with adaptivity and personalization. Therefore, this study is potential for enhancing pedagogical practices in mobile, contextualized, and ubiquitous learning environments in future research.

Acknowledgements

This work is partially funded by JSPS KAKENHI 22H03902, NEDO JPNP18013, JPNP20006.

References


EGG
Relationship Between Students’ Minecraft Re-engagement Metrics and STEM Interest

Maria Mercedes RODRIGO*, Jonathan DL. CASANO & Mikael FUENTES
*Ateneo de Manila University, Philippines
*mrodrigo@ateneo.edu

Abstract: We determine the relationship between student STEM interest and four re-engagement metrics within a set of Minecraft worlds: number of concurrent players playing outside of class hours, number of observations made, number of worlds visited, and number of science tools used. With the exception of number of observations made, there were no significant relationships found between STEM interest and these re-engagement metrics. Therefore, the data so far does not provide much evidence that these behaviors are indicative of interest in STEM. Ways forward include later administration of the STEM interest questionnaire, hypothesizing other in-game behaviors, and identifying a more sensitive STEM interest instrument.

Keywords: WHIMC, Minecraft, re-engagement, Philippines

1. Introduction

The What-If Hypothetical Implementations in Minecraft (WHIMC; https://whimcproject.web.illinois.edu/) are a set of Minecraft-based simulations that immerse students in alternate versions of Earth. They pose questions such as “What if Earth was tilted on its axis?” or “What if Earth had two moons?”. The students then explore these Earths in order to observe how climate, geography, and life forms would differ because of these varying circumstances. A team from the Ateneo de Manila University (ADMU) partnered with WHIMC’s originators at the University of Illinois Urbana Champaign (UIUC) to deploy WHIMC in Philippine classrooms. The goal of the ADMU team was to use WHIMC to cultivate student interest in STEM.

In order to measure the impact of WHIMC on student interest, the work of Casano, Fuentes, and Rodrigo (2023) investigated the concept of re-engagement. A learner is said to be re-engaged if he/she interacts with a learning intervention without a teacher’s prompting. Student engagement takes place when the students use the intervention because he/she is required to do so, during a time that the teacher designates. Casano et al. note that prior studies of student use of Minecraft tends to focus on engagement and sustained engagement, which usually span from the time Minecraft is introduced in a class to the time the students are asked to answer post-intervention questionnaires. While there is some work on student re-engagement with Minecraft, it is not as extensively studied.

Casano et al. (2023) operationalized re-engagement in WHIMC as social play, free exploration, observations made, and science tools used. In this paper, we determine the relationship between these metrics and STEM interest. A student’s social play score was equal to the number of overlapping play sessions, i.e. the number of concurrent companions playing WHIMC with the student outside of teacher-mandated hours. Free exploration was operationalized as the number of worlds the student visited, both during and outside of teacher-mandated hours. Number of observations referred to the count of the observations logged by the student. Finally, the use of science tools referred to the number of times students measured different characteristics of the worlds they visited. The goal of this paper is to determine how these re-engagement metrics relate with student STEM interest.
2. Methodology

The data analyzed in this paper was from eighth-grade students from eight different sections of a private school in the southern part of the Philippines. Their teachers received training on the use of WHIMC, after which they crafted their own modules. These modules underwent a process of review in which curricular alignment and quality were checked. A full discussion of the preparatory process is available in (Manahan & Rodrigo, 2022). After revision, the teachers conducted these lessons with their students.

There were five types of data collected from the students. We collected demographic data about the students’ age, gender, school, and school section. WHIMC automatically logged student interactions with the system. Data from these logs included the student’s position in the world, what observations the student made, and what tools the students used. The students also completed three types of questionnaires. A STEM interest questionnaire (SIQ) was given to students before and after the use of WHIMC. This questionnaire asks students questions about their perceived ability in STEM, their perceived future use of STEM, and their interest and value for STEM. A Game Experience Questionnaire (GEQ; IJsselsteijn & De Kort, 2013) asked students how engaging the game experience was. Finally, Knowledge questionnaires were created by the teachers as part of their lesson planning. Both the GEQ and the Knowledge questionnaires are excluded from the scope of this paper.

There were 175 students who participated in the WHIMC lessons. The dataset for this analysis was composed of the following fields:

- Demographic data: Gender, Age
- Questionnaire data: SIQ Pre-test score, SIQ Post-test score, Difference between SIQ Post-test and Pre-test
- WHIMC log file data: Used WHIMC outside of class (Yes/No), Number of concurrent users during use outside of class, Number of observations made, Number of science tools used, Number of worlds visited outside of class

We removed students with incomplete pre- and post-test data, leaving a total of 131 students for this analysis. Of these 131 students, 58 were male and 73 were female. Their ages ranged from 13 to 15 years old.

3. Analysis and Results

To understand how STEM interest changed and how it related with re-engagement, we first performed a paired t-test to determine whether the SIQ scores of the students changed significantly. While the post-test scores were slightly higher on average (3.6) than the pre-test (3.5), the difference was not significant (t(254.68)=-1.02, p=0.3). This is consistent with Tablatin, Casano, and Rodrigo’s (2023) analysis of a subset of this data, which also showed no significant difference. This implies that using WHIMC did not influence students’ interest in STEM.

There were no statistically-detectable gender differences. Male and female students did not differ significantly in terms of their pre-test scores (t(123.31)=0.63, p=0.5), post-test scores (t(127.14)=-0.56, p=0.58), or their post- and pre-test difference (t(115.38)=1.43, p=0.16).

For the next analyses, we treated the population as a whole and only used the SIQ post-test scores as the outcome variable.

We first attempted to correlate the SIQ post-test scores with the number of concurrent companions, the number of observations, number of science tools used, and the number of worlds visited. All of these correlations were negligible.

We then tried to divide the population into different terciles based on each of the metrics. That is, we first sorted the number of concurrent companions and then divided the population into the top, middle, and bottom third. There should have been 43 to 44 students per tercile. However, if a cluster of students at the border of two terciles had the same scores, we assigned the cluster to either the higher or lower tier, depending on what would result into as balanced a split as possible. We then performed an ANOVA and a Tukey
Honestly Significant Difference post-hoc analysis to determine if there were differences in SIQ post-test scores among the different terciles. There were no significant differences among terciles for the number of concurrent companions (f(2)=1.83, p=0.17), number of science tools used (f(1)=0.73, p=0.4), or number of worlds visited (f(2)=0.36, p=0.72).

For the total observations, there were 75 students with no observations at all and 57 students with at least one observation. Instead of splitting this group into terciles, we split the group into two—those with at least one observation and those with no observations. We found that the students who made no observations had lower SIQ post-test scores than those who made at least one observation (t(128.42)=2.08, p=0.04).

4. Conclusion

The results of the analysis are disappointing. With the exception of number of observations made, none of the re-engagement metrics had a statistically detectable relationship with SIQ scores. The data did not provide evidence in favor of these measures of re-engagement. One possible explanation was that the SIQ post-test was given before the students re-engaged with WHIMC. It is also possible that the exposure to WHIMC was too brief to encourage or inspire greater STEM interest. In future research, we could ask students to complete the SIQ again, several weeks after they complete the WHIMC lesson. We could also search for other in-game behaviors that might be indicative of STEM interest. Finally, we could search for a more sensitive measure of STEM interest, other than the SIQ.

Acknowledgements

We thank H Chad Lane and Jeff Ginger for their enthusiastic collaboration. We thank our partner teachers and schools. We thank Ms. Rosario Madjos for her administrative support. We thank our colleagues at the Ateneo Laboratory for the Learning Sciences. Finally, we thank the Ateneo de Manila University for making this work possible through the grant entitled, “The Impact of Technology Use on Learning Outcomes and the Learning Experience”.

References


Enhancing Learning Experience in University Engineering Classes with Kahoot! Quiz Games

Zilu LIANG*

*aUbiquitous and Personal Computing Lab, Faculty of Engineering, Kyoto University of Advanced Science (KUAS), Japan
*liang.zilu@kuas.ac.jp

Abstract: In this short paper we present a study examining students' experience using a gamification quiz platform, Kahoot!, in two engineering courses in a private Japanese university. We ran Kahoot! quiz game regularly at the end of each lecture session throughout a semester. Students were encouraged but not forced to engage in discussions with peers during gameplay. We evaluated our approach using an original questionnaire designed to probe students' opinions of the usability and usefulness of the quiz games as well as their preferred play mode. The results showed that overall students had positive experience playing the Kahoot! quiz games in class, though many of them had no prior experience using Kahoot! or any kind of online quiz platforms before taking the courses we offered. Surprisingly, their interest in the quiz games did not wear off, and they enjoyed the fun and competitive atmosphere throughout the semester. Consistent with findings in prior studies, our analysis showed that many students considered the quiz games useful in helping them to concentrate in class and to develop a deeper understanding into the technical content. To a less degree, students found the quiz games motivating them to learn more about the subject. Students' preference of the play mode split, which suggests the importance of diversifying the play modes to meet the needs of different types of learners. We reflected over the nature of our quiz questions using the revised version of the Bloom's taxonomy. Our current design was centered on the ‘Remember’, ‘Understand’, and ‘Apply’ levels of factual and conceptual knowledge. Future studies are needed to expand the quiz game design to support the development of higher-order cognitive skills at the ‘Analyze’, ‘Evaluate’ and ‘Create’ levels, as well as adding additional objective measures for the assessment.

Keywords: Kahoot!, gamification, quiz, active learning, engineering education, computing education

1. Introduction

Quiz is widely used as a tool for pre- and post-class assessment (Cavadas et al., 2017; Cook & Babon, 2017; Liang, 2019). In particular, online quiz was found to be effective in incentivizing students’ completion of preparatory work in flipped classroom (Egan et al., 2017), and in engaging and motivating students in remote learning (Parte & Mellado, 2022). Quiz improves knowledge retention and is considered an effective strategy of active learning (Cook & Babon, 2017). However, traditionally quiz is often used as an asynchronous activity where students answer the questions on their own and receive feedback with a delay. Game-based quiz games such as Kahoot! were developed to address the limitations of traditional quizzes and to enhance students learning experience. Kahoot! is a web application that allows educators to create interactive quiz games. It does not require an installation and can be accessed through an Internet browser on any digital device. It has been widely used in classes from K-12 to higher education around the globe (Wang & Tahir, 2020). Kahoot! provides user-friendly graphic interface for instructors to create their quiz activities, and a separate interface for players to answer the quiz items. Kahoot! allows
instructors to add a variety of questions, including single- and multi-choice questions, true or false questions, and puzzles. With a premium subscription, instructors will also be able to embed slides, videos, audios in a quiz game as well as collecting opinions with polls, word cloud and open-ended questions. Allowing for real-time feedback is a main difference between a Kahoot! quiz game and a traditional quiz (either paper-based or digital). For each question, once all students have submitted their responses, the system automatically grades all students and projects the correct answer as well as the distribution of their responses to the whole class. It allows instructors to provide immediate feedback, to comment on common mistakes, and to ask follow-up questions as needed. Such instantaneous interaction is often more educationally effective than when delivered after a delay (Thelwall, 2000). Instructors can customize the time allocated to each question so that the time pressure can be adapted to students’ levels. In addition, by projecting an updated ranking of students after each question, it creates a competitive atmosphere which helps boost students’ motivation and engagement (Wang & Tahir, 2020; Licorish, et al. 2017). At the end of a game, players are ranked based on their final scores and response time. After playing a quiz game, the platform will generate a report summarizing the overall performance of the whole class, and players’ overall scores and scores to each question.

While there is a large body of evidence that the Kahoot! quiz games enhance classroom dynamics and motivate students to engage in the class flow (Cameron & Bizo, 2019; Licorish, et al. 2017; Wang & Tahir, 2020), little is known in terms of whether the quiz games motivate them to learn more about the subject and whether they appreciate the collaborative opportunities enabled by the quiz games. In addition, Asian student populations were under-represented in previous studies. In this paper, we share our experience of using quizzes to enhance learning experience in an English-medium engineering faculty in a private Japanese university. Leveraging the Kahoot! quiz platform, we combined gamification and summative quizzes to create a fun and competitive atmosphere in class, with the main purpose of promoting learner-content interactions and little emphasis on assessment. We evaluated our approach by collecting students’ feedback using a questionnaire. By analyzing their responses, we aimed to generate preliminary answers to the following questions:

- How did students like the Kahoot! quiz games? Did they enjoy competing with peers?
- How often did they engage in discussions with others during gameplay?
- Did students prefer playing individually or in groups?
- Did the quiz games help students concentrate in class?
- Did the quiz games deepen their understanding of the learning content?
- Did the quiz games motivate them to explore more on the subject?

2. Method

We used the Kahoot! quiz platform in two courses. One is a mandatory Android app development course (denoted as C1) designed for first year university students, and the other is an elective digital signal processing course (C2) for third year students. Both courses offered 15 weekly teaching sessions. A teaching session last 3 hours and consist of a lecture or tutorial and some hands-on exercises. Details of the courses can be found in (Liang, 2022; Liang et al., 2021). In the 2022 academic year, 56 and 21 students were enrolled in each course. We played Kahoot! live at the end of each lecture session and alternated between the two play modes: classic mode and group mode. In classic mode students answered the questions with or without peer-interaction and they were ranked individually. In group mode students formed teams and had a short period of group discussion before answering each question. In this play mode students were ranked by teams. In Figure 1 the left pane shows the user interface on students’ device, and the right pane shows the question and options that are projected to the whole class. To avoid overtaxing students’ cognitive resources, our quiz games usually consist of 5 questions that can be completed in less than 10 minutes.
Figure 1. User interface of the Kahoot! quiz game platform.

We developed an original questionnaire to collect students’ opinions of the Kahoot! quiz games. The questionnaire consists of 17 items as shown in Table 1 under the themes of usability, entertainment, competition, reinforcement, and motivation. Students rated each item on a Likert scale ranging from “strongly agree” (=5) to “strongly disagree” (=1). The question items were designed to probe students’ opinions of the usability and usefulness of Kahoot! as well as their preference of collaboration with classmates during gameplay. The questionnaire was implemented using the Microsoft Form.

3. Findings

Table 1. The median and mode of students’ responses for each item.

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Questionnaire Item</th>
<th>Android App Development (C1)</th>
<th>Digital Signal Processing (C2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>It was easy to use the Kahoot! app on my device.</td>
<td>5 (5)</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Q2</td>
<td>It was fun to play the Kahoot! quiz games.</td>
<td>5 (5)</td>
<td>5 (5)</td>
</tr>
<tr>
<td>Q3</td>
<td>It was fun to compete against other students.</td>
<td>4 (5)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q4</td>
<td>I wished to score top 3 and always tried my best to compete.</td>
<td>4 (5)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q5</td>
<td>I was frustrated when I didn’t score well in a Kahoot! game.</td>
<td>4 (4)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q6</td>
<td>I discussed with other students to come up with my answers.</td>
<td>3 (3)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Q7</td>
<td>I prefer playing the group battle mode.</td>
<td>3 (3)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Q8</td>
<td>I was initially excited about Kahoot! in the first classes but gradually lost interest.</td>
<td>2 (2)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Q9</td>
<td>Playing Kahoot! helped me stay concentrated.</td>
<td>4 (3)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q10</td>
<td>Playing Kahoot! deepened my understanding of the learning contents.</td>
<td>4 (4)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q11</td>
<td>Playing Kahoot! motivated me to learn more about the subject.</td>
<td>3 (3)</td>
<td>4 (3)</td>
</tr>
<tr>
<td>Q12</td>
<td>I wished Kahoot! was used in other lectures.</td>
<td>4 (5)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q13</td>
<td>Overall, the Kahoot! quizzes were helpful.</td>
<td>4 (4)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Q14</td>
<td>Overall, I like playing Kahoot! quiz.</td>
<td>4 (5)</td>
<td>4 (4)</td>
</tr>
</tbody>
</table>
We received responses from 47 and 21 students, which translates to a response rate of 83.9% and 100%, in C1 and C2, respectively. As for prior experience with Kahoot!, more than half of the students (N=25) in C1 had no prior experience using Kahoot! before taking the course. Among those 25 students, some had used other quiz platforms such as Quizlet (N=2) and Everyone’s Quick Quiz/ Min’na de hayaoshikuizu (N=6). Similarly, the majority of the students in C2 (N=19) had no prior experience using Kahoot! but a few had used Quizlet (N=2) and Everyone’s Quick Quiz/ Min’na de hayaoshikuizu (N=1). This implies that gamification is not yet widely used in educational institutions prior to university.

Figure 2. Distribution of responses to “Q6: I discussed with other students to come up with my answers” in C1 (Left) and C2 (right).

Figure 3. Distribution of responses to “Q7: I prefer playing the group battle mode” in C1 (Left) and C2 (right).

Given the ordinal nature of the data, we calculated the median and mode (instead of average and standard deviation) of students’ responses for each item, as shown in Table 1. Overall, students’ experience with Kahoot! was positive in both courses. Students considered the quiz games easy to use (Q1), fun to play (Q2, Q14), and helpful (Q13), and many of them wished that Kahoot! was used in other courses (Q12). While we hypothesized that students’ interest in the quiz games would wear out gradually, the result suggested something opposite (Q8). Students in both courses perceived the competitive atmosphere as fun (Q3) and motivating (Q4), though they may feel frustrated if they did not score well (Q5). The opinions split when it comes to collaborating with classmates (Q6, Q7). As shown in Figure 2, 21% and 10% of the students never or seldom discussed with others during gameplay despite of being encouraged to. At the other end of the spectrum, 41% and 47% often or always engaged in discussions with others to come up with answers. As for the
preference of play mode, both courses showed similar tendency. As shown in Figure 3, 33%-41% students preferred group mode over classic mode, 24%-29% students had the opposite opinion, and the rest had no preference. This implies that instructors should alternate between the two play modes to meet the needs of students with different preferences. Initially we feared that the Kahoot! quiz game may be perceived primarily as an entertainment activity but with limited effect on learning because a quiz game was only a brief activity in each class. Nonetheless, in both courses many students agreed that the quiz games deepened their understanding (Q10), and to a less degree, helped them stay concentrated (Q9) and motivated them to explore more about the subject (Q11).

4. Discussions

We have presented our experience of using Kahoot! quiz games in class to engage and motivate students. Our findings echo previous studies that Kahoot! quiz brings the benefit of enjoyment and engagement, which is likely due to the competitive atmosphere created by the point system and the background music (Wang, 2016; Licorish, et al. 2017; Cameron & Bizo, 2019). While previous studies have demonstrated the effect of Kahoot! in motivating students to engage with current learning, our results demonstrate that Kahoot! may also motivate students to explore more of the subject in the near and far future.

Figure 4. The design of our Kahoot! quiz games was centered on the ‘Remember’, ‘Understand’, ‘Apply’ levels of factual and conceptual knowledge.

In what follows we discuss the role of the Kahoot! quiz games through the lens of the revised version of the Bloom’s taxonomy (Krathwohl, 2002). As shown in Figure 4, the revised taxonomy consists of four types of knowledge and six levels of cognitive process organized in hierarchy. The design of our Kahoot! quiz games was centered on supporting students’ learning on the ‘Remember’, ‘Understand’, and ‘Apply’ levels of factual and conceptual knowledge. In particular, the single-choice and true-or-false questions prompted students to recollect the factual and conceptual knowledge covered in the lecture sessions to boost knowledge retention. As indicated by students’ responses to Q10, the quiz games were perceived as effective in helping deepen their understanding of the learning contents, hence supporting the ‘Understand’ level of cognitive process. Given the time pressure for students to answer the questions synchronously, it is challenging to design questions that support the development of higher-order thinking skills at the ‘Analyze’, ‘Evaluate’, and ‘Create’ levels, particularly those related to the procedural and metacognitive knowledge dimensions. Acknowledging the limitations of our quiz game design, we incorporated other activities as a complement. For example, in both courses students were required to work on a self-directed final project towards the end of the semester. The planning, execution,
implementation, and monitoring of the project provided great opportunities for students to acquire procedural and metacognitive knowledge while developing higher order thinking skills. There are several limitations of this work that should be addressed in future studies. First, this study was conducted in a private university and thus may not generalize well to other types of educational institutions. Second, the findings relied on self-reported data from the questionnaire, which may subject to recall bias or social desirability. Objective measures or additional assessment could be used in the future to strengthen the study. Third, we did not delve deeper into the design opportunities for supporting higher-order cognitive skills and collaborative learning using the Kahoot! quiz game platform.

5. Conclusion

We have shared our learning design with the Kahoot! quiz game platform in two engineering courses targeting first- and third-year university students. The quiz games created a fun and competitive atmosphere to engage and motivate students. The questionnaire results demonstrated that students had positive experience with the quiz games. Students found the quiz games useful in helping them concentrate in class and deepening their understanding of the technical contents. Students’ preferences split when it came to whether to engage in discussions with peers and the play mode, which implies the importance of diversifying the play settings to meet the needs of different types of learners. According to the revised version of the Bloom’s taxonomy, our current quiz game design was centered primarily on supporting the development of lower-order cognitive processes including ‘Remember’, ‘Understand’ and ‘Apply’ factual and conceptual knowledge. In our next step we will evaluate our learning design with a large cohort using both subjective and objective measures. We will also expand our quiz design to support the development of higher-order cognitive processes at the ‘Analyze’, ‘Evaluate’ and ‘Create’ levels and collaborative learning.

References

Design and Implementation of an Educational Game for Teaching Artificial Intelligence to High School Students

Ning WANGa, Ryan MONTGOMERYb, Eric GREENWALDb & Maxyn Leitnera

aInstitute for Creative Technologies, University of Southern California, USA
bLawrence Hall of Science, University of California, Berkeley, USA
*nwang@ict.usc.edu

Abstract: Artificial intelligence (AI) is becoming ubiquitous in our daily lives. In more and more fields, AI systems are transforming how knowledge is constructed, discoveries are realized, and how solutions are developed and tested. These changes have profound implications for the future workforce and citizenry. Yet, learning AI remains a niche subject largely reserved for advanced post-secondary educational contexts. While there is growing attention to broadening AI educational opportunities and, especially, to providing learning experiences for younger students, relatively little is currently known about how to most effectively provide AI education to K-12 (kindergarten through 12th grade) students. In this paper, we discuss the design and present findings from an implementation study of an educational game for high-school AI education called ARIN-561. Drawing on an integrated analysis of gameplay log data, pre/post knowledge, and disposition surveys for nearly 1,000 high school students, we present findings on the efficacy of the educational game and its constituent activities in advancing AI learning goals. We explore possible interactions between learning outcomes, incoming math knowledge, prior gaming experience, and other factors that can inform future learning design and shed light on what can position youth for success in game-based AI learning experiences.

Keywords: K-12 AI education, youth AI education, educational game

1. Introduction

Artificial Intelligence (AI) is a foundational technology permeating every aspect of our daily lives. Rapid advances in the design and implementation of AI systems have led to the ever-expanding role for AI in society. It is also profoundly transforming our workforce around the globe. While some of today’s youth will become the future AI workforce and a majority of them will join a workforce that utilizes AI, all will become end-users, such as consumers of AI (Gardner-McCune et al., 2019). It is critical, therefore, to prepare future generations with basic knowledge of AI, not just through higher education, but beginning with childhood learning.

While AI’s impact on society is deepening and expanding in myriad ways, and innovative educational opportunities are being rapidly developed, there has been little research into how students, especially pre-college students, construct an understanding of and gain practice with core ideas in the field. As a result, there is yet little possibility of grounding the design of learning experiences in evidence-based accounts of how youth learn AI concepts, how understanding progresses across concepts, or what concepts are most appropriate for what age-levels. AI is built on a foundation of philosophy, psychology, and mathematics, and it centers around using algorithms to solve real-world problems (Russell and Norvig, 2016). This provides a theoretical foundation to connect AI learning to existing Science, Technology, Engineering, and Mathematics (STEM) subjects in K-12 classrooms. Given the packed schedule of existing courses of K-12 students, it becomes a more realistic approach
to embed AI education in K-12 classrooms. Such an approach to AI instruction offers a rich context to learn scientific and mathematical concepts already taught in K-12 (Wang and Johnson, 2019) and to apply them to problem solving.

One technology-based approach to bring AI to the K12 classroom that has shown promise in other STEM disciplines is digital game-based learning. Decades of research evidence point to the efficacy of game-based learning in promoting student learning (Plass et al., 2020). However, there is little research into using game-based learning for AI education for youth (Lee et al., 2021), given that the research field of K-12 AI education is still in its infancy. In this paper, we will discuss the design and initial implementation study of an educational game, called ARIN-561, for teaching high-school students about AI. We conducted an evaluation study at high schools in the United States. Results indicate the potential of ARIN-561 to build AI knowledge, especially for students who have background knowledge with the relevant mathematical concepts typically taught at the high school level.

2. Related Work

AI education has long been absent from K-12 classrooms. Recent efforts are beginning to investigate the integration of AI into K-12 schools, including defining AI literacy (Long and Magerko, 2020) and developing curricula and guidelines (Gardner-McCune et al., 2019; MIT AI Education Initiative, 2021). Researchers in youth AI education have been experimenting with teaching AI, including machine learning (Rodríguez-García et al., 2021; Zhou et al., 2021) and ethics (Forstyh et al., 2021), within the context of computational thinking (Ritter et al., 2019) through conversational agents (Lin et al., 2020), dance (Payne et al., 2021), and game-based learning (Lee et al., 2021). Discussions on youth AI education are heating up in Europe (Kandlhofer et al., 2019; AI+, 2021), China (Peterson et al., 2021), Israel (Shamir and Levin, 2020), and around the world (Youjun et al., 2018; Yukun and Tang, 2018). For example, researchers in Thailand have designed an agricultural-based AI challenge to foster middle-school students’ learning of the machine learning process in the form of a game (Sakulkueakulsuk et al., 2018), where students build machine learning models to classify ripe or unripe mangoes. In Australia, researchers have designed and implemented classroom activities for teaching fundamental concepts of AI to Year 6 students to demystify AI through activities such as an unplugged activity on facial recognition and a simple robotic exercise that introduces the concept of machine learning (Ho et al., 2019).

The work presented here aims to uncover how to design an educational game to meet the challenges of teaching AI to K-12 students. This work builds upon explorations into how K-12 students approach AI concepts, what obstacles they face, and how to guide them through obstacles (Greenwald et al., 2021). This work also draws upon previous investigations into linking AI to the K-12 math curriculum to identify AI concepts suitable for high school students (Wang and Johnson, 2019), as well as work investigating the learning of computational thinking (Lee et al., 2011) and seminal research into comprehension of mathematical representations (e.g., Curcio, 1987; Friel et al., 2001).

3. ARIN-561 Game-Based Learning Environment

ARIN-561 is a 3D role-playing game designed to teach high-school students AI concepts, prompt them to apply their math knowledge, and develop their AI problem-solving skills. In the game, students play as a space-faring scientist who has crash landed on an alien planet, named ARIN-561 (Figure 1). In order to safely return home, the scientist begins exploring the planet to gather resources needed to repair the broken ship while uncovering the mystery of the planet. The activities for survival and for exploration form the basis for the tasks the students carry out in the game. The game currently covers three classical search algorithms: breadth-first search (BFS), depth-first search (DFS), and greedy search. Each topic consists of two modules: a tutorial module (e.g., Figure 1 bottom left) and a transfer module (e.g., Figure 1 bottom right). Embedded in all the tutorial and transfer modules are
quizzes that help students pause and self-assess (Figure 1 top right). In-game challenges, such as searching for missing spaceship parts or cracking passwords, serve as natural opportunities for the introduction of search as a topic. The essential concepts such as space and time complexity also lend opportunities to connect math knowledge familiar to high school students and these AI concepts that are usually taught in higher education. The integrated educational content in ARIN-561 leverages this opportunity by supporting the students’ application of math knowledge to the evaluation of each algorithm as they progress through the game. In addition to the learning modules, students can also explore the game environment for “off-task” activities (Sabourin et al., 2011), such as gathering minerals around the planet.

![Figure 1. Screen capture from ARIN-561. Top-left: The player crash landed on a foreign planet. Top-right: student is presented with a quiz question about estimating the complexity of search algorithms. Bottom left: student think-alouds through the greedy search algorithm. Bottom-right: the student solves an 8-puzzle using one of the search algorithms to fix their companion robot’s circuit board.](image)

3.1 Learning Experience Design in ARIN-561

The design of ARIN-561 is guided by lessons learned from pilot studies on student AI problem-solving (Greenwald et al., 2021). We also developed additional design principles based on our observations of the characteristics unique to AI learning, and how lessons learned from related fields, such as computer science education (Lee et al., 2011), could be used to inform the design here.

**Facilitate Abstraction** In a cognitive interview study, researchers explored how K-12 students, particularly those in high school, approach AI concepts, what obstacles they face, and how to guide them through obstacles (Greenwald et al., 2021). In the study, students were presented a set of AI problems in a wide range of topics, such as various machine learning algorithms. Students’ think-alouds as they attempted to solve the AI problems (using paper and pencil) shed light on the critical step most students struggled with — problem formulation, or the Abstraction phase in computational thinking, i.e., **Abstraction, Automation, and Analysis** (Lee et al., 2011). This is an initial step where students formulate a problem described in natural language (e.g., find the shortest path) into one that can be...
solved by a computer, such as creating variables (e.g., distance), determining end conditions (e.g., search ends when distance can’t be minimized). Strategies employed by the expert interviewer provided a basis for pedagogical design in ARIN-561.

One of the design decisions made to facilitate Abstraction was to display the real-world problems and the abstract representations side-by-side and to update both synchronously (e.g., the lower-left screen in Figure 1). In an ARIN-561 route-planning problem where students are tasked to use search algorithms to plan a route to a waterfall, a map is placed on one side of the screen while a search tree is displayed over and next to the map. As students direct the algorithm by exploring locations, connected via roads, on the map, the search tree updates accordingly step-by-step and illustrates how locations are represented as nodes and roads are represented as edges in a tree data structure, and how route-planning on a physical map is computationally solved as the expansion of a search tree. When a goal node is reached in the search tree, it is highlighted both on the search tree as a path across edges from root to the goal node, and on the map as a route reaching the waterfall connected via roads.

**Learning Transfer** In education, transfer of learning occurs when learning in one context enhances or undermines a related performance in another context (Perkins et al., 1992). During learning transfer, students apply learning in one discipline across multiple situations. Transfer of learning is particularly important for AI education, as AI can be considered as a discipline of using algorithms to solve real-world problems. When students learn how an AI algorithm can be used to solve illustrative problems in one domain, it is critical to also guide them through problems from a different context to help them build the connection — the abstract representation of the algorithm that can be applied to formulate solutions to seemingly different and unrelated problems.

For each algorithm covered in ARIN-561, we developed a tutorial problem and a transfer problem. The two sets of problems are different enough to arguably be considered as far transfers (instead of near transfers) (Perkins et al., 1992). Tutorial problems are chosen from domains familiar to the students, such as finding a route from point A to B on a map. In a typical tutorial module, students are scaffolded through the abstraction, automation, and analysis processes (Lee et al., 2011) through the player character’s think-aloud and their dialogue with the companion robot. In the abstraction phase, the students are guided to create an abstract representation of the practical problem. After students demonstrate their understanding by correctly expanding the tree for several levels, they are provided with the option to automate the process. In the automation phase, students can watch the search tree continue to expand automatically on the same interface — physical map and abstract search tree placed side-by-side. Students can also pause and step through the tree expansion one step at a time to examine the process closely. The automated expansion animation helps illustrate the characteristics of the search algorithms, e.g., expanding in a breadth-first or depth-first fashion. In the analysis phase, students are guided by the game narrative to examine the solution (e.g., the route found) and to evaluate the process through which the solution was generated (e.g., time and space complexity of the search algorithm). The subsequent transfer problem module presented students with a different problem, such as cracking a password or solving an 8- puzzle. Students were guided by similar but much abbreviated scaffolds through the Abstraction, Automation, and Analysis processes in the transfer phase.

**Comparative Explanation** AI is human ideas represented mathematically and realized computationally. From Classical Search to Local Search, from Propositional to First-Order Logic, from Decision Trees to Genetic Algorithms, AI algorithms build on each other: a new algorithm is often created by modifying an existing one, to solve problems that the existing one was not suited or able to solve. This insight creates both challenges and opportunities for AI education. The evolutionary characteristics of AI algorithms provide a basis for pedagogy that leverages students’ prior knowledge (of an algorithm they are already familiar
with) while constructing the new ones. By directly comparing the new and old algorithms, for example, students not only learn the new, but also reinforce the learning of the old. Such comparisons are not just algorithmic, but also the contextual in terms of application. Understanding the pros of the new and cons of the old in what problems they are or are not suited to address is a key to using AI for problem solving. The approach of prior knowledge activation is not new (Alvermann et al., 1985), nor is the issue of activating inaccurate prior knowledge (van Loon et al., 2013). The explicit and direct comparison between the new and old, when discussing the new, may offer an opportunity for students to reexamine their misconceptions of the old.

In ARIN-561, game modules are organized by learning topics, such as BFS and DFS. After scaffolding students through the first AI algorithm (such as BFS), each new AI algorithm (e.g., DFS) is introduced through an example problem that the previous algorithms fail to solve (e.g., computer runs out of memory when using BFS for route planning). The students are then guided through the Analysis phase to uncover why the previous algorithm failed (e.g., storing too many nodes in computer memory) and how to modify it to address its weakness (e.g., prioritizing expanding child nodes instead of sibling nodes in the search tree). Such modification thus results in the birth of the new algorithm (e.g., DFS). The direct comparisons are not only realized in the explanations through the game narrative, but also illustrated on the user interfaces across the learning of different algorithms.

4. Methods

4.1 Recruitment

Twenty-three math, science, and computer science teachers from a school district in a major metropolitan area in the United States participated in the study. 1274 high school-aged youth from classes taught by participating teachers were recruited for the study.

4.2 Procedure

Participating teachers were provided an overview of the game, learning goals, and study procedure a few months before the study began. A few weeks prior to the study, students were given an online parental consent form and a youth assent form. Only students who consented participated in the study. The study was carried out over 4 class sessions, each lasting 45-55 minutes long, with at least 2 class sessions dedicated to individual gameplay for students. During the first session, students were first assigned IDs to protect their identity throughout the study, and then completed the pre-survey online. At the end of the first session, students logged into the ARIN-561 game online via a web browser. Any technical difficulties encountered were addressed during the first session, via support from the research team. During the second and third sessions, students continued to interact with ARIN-561 at their own pace. Game progression, play time, and answers to in-game questions were recorded for each participant. During the fourth session, students completed the post-survey online.

With restricted access to school campuses due to COVID-19, the study was carried out entirely by the participating teachers. The research team did not participate in the data collection. Additionally, because students were not required to answer all the questions on the pre- and post-surveys, there are missing data at the item level for some students.

4.3 Measures

The pre-survey consisted of items about students’ demographic background, AI Use Type, Interest in AI, AI Knowledge (15 questions), Math Self-efficacy (Liu and Koirala, 2009), and Math Knowledge. All scales except the Math Self-efficacy were developed by the research
team. The AI Use Type included items such as “When I think about how I’d like to interact with AI in the future, I expect that: I will use AI systems in my everyday life as a consumer, and I expect to USE AI systems as a part of my job.” The Interest in AI scale included questions such as “Outside of school I try to learn a lot about AI.” The assessment of AI knowledge and math knowledge specifically focused on the content covered in ARIN-561, in the format of multiple-choice questions. The AI questions were set in the context of solving AI problems similar to those encountered in the game. The questions assessed students’ understanding of, for example, pros and cons of the search algorithms, search algorithms most applicable to specific types of problems, etc. In the post-survey, the same items on interest in AI and AI knowledge from the pre-survey were included. In addition to the surveys, game logs from ARIN-561 were collected. The logs included the in-game click-stream data and responses to in-game quizzes.

5. Results

Of the 1274 participating students, 1014 completed the post-survey. The research team was able to match pre-, post- surveys, and game logs for 764 students. Other than normal attrition (e.g., students absent at either pre, post administration, or game play class), additional data loss was primarily due to errors in student ID entries on the survey platform, which resulted in mismatches of student IDs between both surveys and game logs. We conducted ANOVA analyses to ensure the final sample of 764 students was not significantly different from the full participant sample in terms of background, such as gender, race/ethnicity, and prior mathematical knowledge.

The participants’ average age was 16, with 18% 12th graders, 30% 11th graders, 23% 10th graders, and 29% 9th graders. A total of 46% of the students identified as male, 48% identified as female and 6% identified as other categories or preferred not to disclose. 27% of the students speak English at home, 67% speak both English and a second language at home, and 6% speak only a language other than English at home. Spanish is reported as the non-English language for those students. Interestingly, even though ARIN-561 and the surveys are offered in both English and Spanish, and the teachers were briefed about the language choice prior to the study, all the students chose to use the English version of the surveys and the game.

5.1 AI Learning Gain

We conducted a paired sample t-test on the AI knowledge scale from pre- and post- surveys to examine if playing the game resulted in gains in AI knowledge. Table 1 summarizes the pre/post changes in AI Knowledge and in sub-constructs directly relevant to modules in the game (additional AI knowledge items covered in broader topics such as search tree representations). Results show that students who participated in the study demonstrated statistically significant gain in AI knowledge, with a mean difference of 0.37 on a 34-point scale ($p = .011$). The AI knowledge scale include 3 sub-scales for each of the search algorithms covered in the game (BFS, DFS, and Greedy search). Additional paired-sample t-tests revealed a statistically significant gain for BFS learning (mean difference of .30* on an 11-point scale, $p = .001$), a smaller and not statistically significant change for DFS learning ($p = .088$), and a nearly flat outcome for the items focused on the Greedy search algorithm.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Max-Score</th>
<th>T-statistics</th>
<th>p-value</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>14.16</td>
<td>14.53</td>
<td>34</td>
<td>2.54</td>
<td>0.011</td>
<td>0.105*</td>
</tr>
<tr>
<td>BFS</td>
<td>5.30</td>
<td>5.60</td>
<td>11</td>
<td>3.25</td>
<td>0.001</td>
<td>0.149*</td>
</tr>
<tr>
<td>DFS</td>
<td>4.36</td>
<td>4.50</td>
<td>9</td>
<td>1.71</td>
<td>0.088</td>
<td>0.081</td>
</tr>
<tr>
<td>Greedy</td>
<td>2.09</td>
<td>2.06</td>
<td>3</td>
<td>-0.48</td>
<td>0.629</td>
<td>-0.022</td>
</tr>
</tbody>
</table>
5.2 Student Background and AI Learning

In the pre-survey, we gathered data on students’ demographic background, such as gender, grade level, language spoken at home, and video game experiences. ANOVA tests show that pre/post AI learning gains did not differ significantly between students of different gender, grade-level, and language spoken at home. Learning gains differed however between students with different prior gaming experience (Figure 2). The participants reported a wide range of gaming experiences. Given the detailed categorization of gaming experiences, we grouped the students into two groups: those who don’t play video games or play 1-2 hours per week (60% of the sample), and those who play 3 or more hours per week (40%). Students who play video games less than 2 hours per week had significantly lower gain on overall AI knowledge (M<2h = .0132, M>3h = .8428, p = .006), including sub-scales on BFS learning (M<2h = .0132, M>3h = .8428, p = .003), Greedy learning (M<2h = -.18, M>3h = .11, p = .019) but not DFS learning (M<2h = .04, M>3h = .29, p = .152), compared to students who play video games 3 or more hours per week.

The pre-survey also includes items that measure Math Self-Efficacy, (relevant) Math Knowledge, and Interest in AI. We conducted a series of regression analyses to investigate these three student level factors that may be predictive of observed learning gains. We found that the prior Math Knowledge (as demonstrated on the pre-survey item set) predicted observed AI learning gains (R = .1, p = .006). This suggests that relevant math knowledge is weakly but significantly related to higher AI knowledge gained through ARIN-561.

![Figure 2](image)

*Figure 2. Left: Percentage of students with different weekly gaming experience. Right: AI Knowledge gain broken down by weekly gaming experience.*

5.3 In-Game Progress and AI Learning

The pedagogical design of ARIN-561 is based on the hypothesis that AI algorithms build on each other. Algorithms, such as DFS introduced later in the game are discussed in comparison to previously introduced algorithms, such as BFS. While students can jump through different modules by going through the menu selection screen in the game, overall, students took a relatively linear path through the game, by going through BFS, DFS, then Greedy game modules. Thus, as students progress through the game, mastering previously discussed algorithm should help students’ learning of the new ones, while learning the new algorithm helps student reinforce the learning of the older ones. We analyzed how reaching milestones in the game, such as completing the DFS module (both tutorial and transfer problem modules), impacts overall AI learning and the learning of individual algorithms. Independent sample t-test shows there is no significant difference in AI knowledge gain between students who completed all modules of the game and students who did not (N< = 556, N> = 208, T: completing, F: not completing, p = .642). Students who completed the BFS module did not gain significantly more AI knowledge overall (F = 1.169, p = .28) or BFS knowledge (F = .5, p = .48) than those who started but didn’t complete BFS modules. However, completing BFS did help students gain more knowledge on DFS (F = 4.545, p =
.033) and greedy search (F = 3.204, p = .074). Completing the DFS or greedy modules did not have a significant impact on overall or individual AI algorithm learning. Given that all students are given the same amount of time to play the game in the classroom, we did not analyze how time in game impacted AI learning. Overall, students spent between 3 seconds and 338 minutes in the game, with a mean/median gameplay time of 89/84 minutes. The outlier of extremely long game-play time is likely due to students forgetting to log out of the game at the end of the class.

6. Discussion

This study demonstrates that a relatively brief in-school exposure to AI learning experiences, via an educational game, can result in learning gains for AI content with pre-college aged youth. Examination of the subscale scores for the AI Knowledge assessment indicated that the learning that took place was concentrated on the BFS algorithm, with smaller gains for DFS, and no gains for Greedy. Drawing on the learning and assessment design concept of a learning progression (Wilson, 2009; Duncan and Hmelo-Silver, 2009), our conjecture is that the design of gameplay, in which students first encountered BFS then compared it to each of the next two algorithms as the game progressed (DFS, then Greedy), led to consolidation of understanding related to BFS and thus a deeper opportunity to learn that content. However, we did not observe a statistically significant impact of completing the BFS, DFS, or Greedy modules on BFS learning gains. Completing the BFS modules however, did contribute to learning DFS and Greedy search. This suggests the efficacy of the progressive roll out of content in our design, where each new content area is explicitly related to prior content. Later design iterations will look to extend opportunities for students to connect and consolidate their emerging understanding of content encountered later in the game, e.g., through additional integrative activities.

The significance of prior mathematical knowledge for predicting observed AI learning gains suggests an educational game that is optimized for youth who already enter with a strong mathematical foundation. This would challenge efforts at using the current iteration of the game for a broad high-school population with a wide range of prior math competencies. Future design iterations will look to support students with varying levels of prior mathematical knowledge, either through focused tutorials for related math content and/or through improved game design that better resonates with students who have not yet taken advanced math courses.

We also see promise for this game-based instructional model in the feasibility of its implementation. First, the youth who engaged with the game did so largely independently of a teacher. The minimal need for outside expertise means that implementation is likely to be less dependent on having educators with AI and computer science expertise, an important consideration given widely reported shortages of high-school teachers with such expertise. Also related to feasibility of implementation, completing the game took roughly 2 class periods on average, which minimizes the time it may draw away from existing scope and sequence. Thus, the educational game is well-positioned to be integrated into a wide range of courses and instructional contexts. On the other hand, the AI knowledge growth was relatively small (effect size = .105), indicating that spending 2 classes playing an AI-themed video game is unlikely to contribute to learning gains that might be expected from a fuller instructional sequence or dedicated course. Interestingly, during post-implementation conversations, the participating teachers expressed strong interests to integrate classroom discussions with game-based learning, and suggested dividing the classroom time into independent gameplay and post-gameplay whole-class discussion, where teachers organize discussions to help students reflect upon what’s learned through the game. Such an integrated approach has the potential to further enhance the efficacy of the educational game.
The study was dependent on a researcher-developed measure of AI knowledge, with limited evidence available of its validity with the population sampled. This speaks to the current dearth of AI knowledge measures developed for precollege-aged youth, a challenge that our research team, and others, are working to address through ongoing research and measurement development. In this specific case, we note that the assessment was likely too difficult for the sampled population (for example, the mean score on the post-intervention administration was 14.53 out of a possible 34 points), limiting its potential for demonstrating the learning of high-school-aged youth. Additionally, the measure included different numbers of items for each of these subscales, with fewer items for DFS and Greedy compared with BFS. This constricted the available range for movement on those constructs, potentially impacting observability of changes that may have taken place. Our team is conducting psychometric tests (classical and IRT methods) and triangulating that with in-game opportunities to demonstrate understanding to inform revisions to our instrument for this population.

Acknowledgements

This research was supported by the United States National Science Foundation under Grant #1842385. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


Executive Functions Training-oriented Digital Games: Effectiveness and Experience

Chi-Fang HUANG\textsuperscript{a*}, Hui-Ling HSU\textsuperscript{a} & Zhi-Hong CHEN\textsuperscript{a, b}

\textsuperscript{a}Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taiwan
\textsuperscript{b}Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan
*81008001e@ntnu.edu.tw

Abstract: Executive function governs important cognitive abilities such as cognitive resource allocation, which is related to daily life and can affect learning performance. However, current executive function training mostly relies on self-administered tests, which leads to a dull process, and low interest, thereby influencing training effectiveness. To assist the affective domain during training, previous research attempted to gamify the training. However, prior studies indicated that the training intention might be influenced by the game-based design, and the game scenario also impacts the training process. Therefore, we developed a training-oriented digital game that closely simulates real-life situations to train executive function based on classic training, and discussed its effectiveness, and user experience. 38 college students were recruited to experience the training-oriented digital game and classic games and found that the training-oriented digital game effectively improved inhibition ability in executive function. Additionally, the game developed in this study showed significantly higher confidence and satisfaction than the classic game, while there was no significant difference in flow, positive affect, and behavioral intention. This indicates that the game developed in this study could be used to train executive function while maintaining support for the affective domain of gaming. We discuss further implications in the study.

Keywords: Executive functions, Digital games, Situational, Training-oriented game

1. Introduction

In the field of cognitive development, the development of executive function (EF) and related research have received increasing attention. EF refers to the cognitive ability to regulate and allocate cognitive resources, which significantly impacts reading, problem-solving, planning, decision-making, and other life-related skills (Miyake et al., 2000; Baggetta & Alexander, 2016). In comparison to Intelligence Quotient (IQ), which is commonly used in intelligence testing and future performance prediction, EF is a better predictor of life quality, mental and physical health, and future achievement (Ahmed et al., 2019; Follmer, 2018; Moffitt et al., 2011). This shows the importance and recent trend of studies related to EF. Miyake et al. (2000) identified three core cognitive abilities of EF: updating, inhibition, and shifting. Among these three, inhibition refers to the ability to resist interference from irrelevant information or impulses during cognitive processing, emphasizing the control of cognitive resources. As human attention is limited, inhibition has a certain degree of influence on other abilities and performances. This ability not only requires control over cognitive resources, but also affects other cognitive abilities or performances related to everyday situations, directly or indirectly affecting learning outcomes, work, and life. Therefore, it is particularly important and deserves special attention.

Studies have found that EF is not only an innate ability but can also be improved through training (Diamond, 2013; Kassai et al., 2019). And digitalized training has been particularly effective in enhancing EF (Jak et al., 2013). However, previous studies often use existing EF tests with minor adjustments and repeat them to achieve training effects, such as
n-back (Soveri et al., 2017; Lo et al., 2021), Go-NoGo (Tschuemperlin et al., 2019; Najberg et al., 2021). Therefore, these training systems are often composed of simple materials and continuous repeat of training. This leads to boredom and reduced intention for training (Prins et al., 2013). Moreover, the training effects may decrease over time without continued training, resulting in a need for longer training periods or repeated training sessions to maintain training effects (Anguera et al., 2021; Söderqvist et al., 2012). Therefore, to maintain and sustain the training effects of EF and to affect daily life positively, it is crucial to consider users’ affective experiences and intention of training.

To improve user experience and training intention, digital games are often used in the training process. For example, McCord et al. (2020) asked elderly people to play Star Wars Battlefront and show significant enhancement on EF after three weeks. Both short-term and long-term training is effective in achieving training goals and positively impacting affective aspects, optimizing the gaming experience, and enhancing training intention (Anastasiadis et al., 2018; Stanmore et al., 2017; Bediou et al., 2018; Nagle et al., 2018). Martinecovic and Vranic (2020) also used multiple games for EF training. However, no significant improvement was shown and this may be because the casual video games are not designed for EF training. Despite the effectiveness of enhancement on EF divided in prior studies, the positive effect of game-based design is still undeniable. Digital games provide adaptive information and training content based on users’ needs, levels, and performance (Bennis & Amali, 2019). The level settings in games not only allow users to choose suitable levels for practice, but also challenge users to establish goals, enhance achievement, stimulate learning motivation, and even affect metacognition and performance (Sun-Lin & Chiou, 2017).

In recent years, there have been studies attempting to develop game-based training to especially aim at improving EF. Homer et al. (2019) focused on training inhibition ability in nearly a hundred adolescents, using digital games to improve EF. They found that not only the training significantly improved EF, but the design of game elements also impacted users’ preferences and interests. In other studies, Ober et al. (2021) developed a digital EF training game called Gwakkamole, which required participants to click on regular avocados on the screen and avoid clicking on avocados wearing a helmet. The empirical results indicated that this game effectively improves EF. Another study found that the effectiveness of different sub-abilities of EF varies in the same training. Moreover, the design of challenging game elements in the game also has an impact on user experience (Wells et al., 2021). In addition, there is a series of studies that developed EF training games with aliens, including The Alien Game (Homer et al., 2018) and All You Can ET (Mayer et al., 2019). The empirical studies using these games to train EF, show significant training effectiveness of EF. Aside from the effectiveness, the impact of game design on learners is also discussed. In addition to establishing goals and achievements in challenging levels, which can help improve training effectiveness, real-time feedback in the game also has certain assist effectiveness for users (Parong et al., 2020).

Notably, most scenario designs are set in virtual worlds, far from real-life situations. However, Gray et al. (2019) have suggested that the effectiveness of EF training is influenced by the degree to which the training process mimics real-life experiences and actions, and whether the effects can be maintained. Therefore, this study aims to develop a training-oriented digital game that is closer to real-life situations to explore whether it can achieve the goal of improving EF and to further discuss the gaming experience when training. The research questions are as follows: 1. Does the training-oriented digital game enhance inhibition ability more effectively than classic games? 2. Does the training-oriented digital game produce a more positive gaming experience than classic games?

2. System design
This study aims to develop a series of 4 EF training games based on real-life scenarios, with inhibition as the training target, as shown in Figure 1. Based on the commonly used Simon task (Simon et al., 1967) and Stroop test (Stroop, 1935) for measuring inhibition ability, the game system includes two types of games: button selection and voice recognition. The button selection games are mainly adapted from the Simon task and are divided into three stages with increasing difficulty. In the first stage, users need to click on the left or right according to
the required rule after viewing an image. After getting familiar with the rules, the second stage of the game introduces interference from the left and right movements and flips to train users' ability to inhibit habitual responses. The third stage adds rule changes, requiring users to make choices opposite to their familiar operations in specific situations, in order to train rule switching, inhibition of previous habits, and increase game challenges. On the other hand, the voice recognition game is based on the Stroop test, and in multiple consecutive stages, users are required to speak out the content according to the rule requirements. During this period, rules are gradually stacked, and users need to inhibit interference from other types of icons and speak out corresponding content based on the rules.

The system was developed using Construct 3 with the aim of training EF skills. To enhance the user gaming experience and encourage intention to keep on training, game elements are added, which include real-life scenarios, sound effects, animations, levels, time limits, and scores. Real-life scenarios are the main characteristics of the game design to make the training process more realistic, and we selected sports, food, and pet interaction as the theme of the game series. Different training scenarios, such as games, restaurants, and rooms, were presented in rich colors to enhance the sense of immersion. Aside from the scenarios, sound effects and animations were provided to offer users instant feedback during transitions, enabling them to know immediately whether their answer was correct or incorrect. This helped users to stay focused on the training objectives and remain immersed in the process. Then, the levels were designed to gradually increase in difficulty. This adds a challenging element to the game. Also, time limits were incorporated into the game to gain challenges. A countdown timer is displayed on the screen to remind users to respond quickly and accurately, with 10 seconds for each response, and faster responses result in higher scores. Last, the scoring system is also important. A scoring board or heart icons that closely matched the packaging scenarios, providing users with a quantifiable and intuitive sense of achievement.

The research design focused on examining the effectiveness and experience of training-oriented digital games developed in this study. The game usage was set as the independent variable, with the participants' original EF ability as the control variable, and the training effectiveness and experience as the dependent variables. To address the research questions, a pretest was conducted before the experiment to assess the participants' initial EF ability. After each intervention using the game, a post-test was conducted to examine the training effectiveness, and a questionnaire was administered to understand the game experience, as shown in Figure 2.
For the game intervention, the participants were divided into an experimental group and a control group, which used the training-oriented digital game and classic game, respectively. The study design referenced previous research designs (Homer et al., 2019; Plass et al., 2019), and the participants were instructed to operate the digital game for 15 minutes within an effective training period that did not induce fatigue.

3.2 Participants

This study recruited 38 college students without any cognitive-related disorders as participants, aged between 20 to 24 years old, including 10 males and 28 females. To avoid the influence of gaming experience on the training process and experience, the weekly gaming duration of the participants was investigated. 18 participants played digital games for more than half of the days in a week, while 20 participants played games for less than half of the time. This shows an evenly distributed pattern and is expected not to affect the research results.

3.3 Research instrument

The research instruments used in this study include not only the training-oriented digital games mentioned above but also classic games used in the control group as a comparison. To assess the effectiveness of the intervention, both EF assessment and questionnaires were used to examine both cognitive and affective aspects, which are described in detail below.

To parallel the training-oriented digital games, 4 classic games were selected. The selected classic games were common ones, including Pong, FreeCell, Pinball, and Minesweeper. These games were chosen to familiarize participants with the game rules. And to examine the effectiveness of the training, Trail making test was used in the pretest and post-test to assess participants' original inhibition ability and training outcomes. The test is frequently used in previous studies related to EF training (Sousa et al., 2020; Sosa & Lagana, 2019). Comprises a series of circle-linking tasks, and participants have to inhibit the effects of different types of circles, while others require participants to link circles in a specific order. After the intervention, a five-point Likert scale questionnaire was used to understand participants' gaming experience. The gaming experience questionnaire includes three dimensions, namely flow, positive affect, and behavioral intention (Law et al., 2018; Hwang et al., 2013). The flow and positive affect dimensions consist of 4 questions each, while behavioral intention has 5 questions.

3.4 Data analysis

To answer the research question, this study utilized IBM SPSS for data analysis. Specifically, an Analysis of Covariance (ANCOVA) was conducted using the differences in games as the between-group variable, pretest scores as the covariate, and post-test scores as the dependent variable to examine whether there were significant differences in performance between groups after intervention. In addition, an Analysis of Variance (ANOVA) was employed to investigate the user experience, with game intervention differences as the independent variable and participants’ experiences as the dependent variable. The questionnaire on experience was used for comparative analysis.

4. Result and discussion

4.1 Training effectiveness

According to the research results, there is a significant difference between pre-test and post-tests of inhibition ability, as shown in Table 1. Through ANCOVA, it shows no significant difference in pre-test scores between the two groups. Moreover, after using the training games, the inhibition ability of the training game group was significantly improved compared to the classical game group, with an average increase from 0.37 to 0.51, and an F-value of 5.34. It can be concluded that participants were able to enhance their inhibition ability more effectively after using training-oriented digital games than classical games.
Table 1. The result of training effectiveness on inhibition

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-test Mean</th>
<th>Pre-test SD</th>
<th>Post-test Mean</th>
<th>Post-test SD</th>
<th>F</th>
<th>etat²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhibition</td>
<td>0.37</td>
<td>.29</td>
<td>0.51</td>
<td>.24</td>
<td>5.34*</td>
<td>.132</td>
</tr>
<tr>
<td>Training game (n=19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classic game (n=19)</td>
<td>0.69</td>
<td>.29</td>
<td>0.39</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05 ; **p < .01

The results of the present study are consistent with previous research, which showed that EF training-oriented games can effectively achieve the goal of enhancing EF abilities. Homer et al. (2019) conducted inhibition ability training on adolescents using a digital game and found significant improvement in training effectiveness. Similarly, Wells et al. (2021) developed a digital EF training game, Gwakkamole, and not only indicates that this game could effectively train EF but also mentions that the design of challenging game elements in the game would also have an impact on the training process. Another series of studies of EF training games take aliens as the game theme, including The Alien Game (Homer et al., 2018) and All You Can ET (Mayer et al., 2019). Empirical research similarly found significant improvements in inhibition ability and noted the impact of game design on learners. Establishing goals and achievements through the challenges of the game can help improve training effectiveness, and real-time feedback in the game can also assist users (Parong et al., 2020). These findings are consistent with the present study, as the design of challenging elements, such as levels and scores, as well as real-time feedback, such as sound effects and animations, has become an essential part of the game process to assist learners in achieving better training results.

However, the above research scenarios differ from real-life situations. The present study takes activities that are closer to real-life situations as the game scenario, based on the gap identified in previous research (Gray et al., 2019). The results of the study also showed that designs that are closer to real-life situations can effectively improve inhibition ability.

4.2 Game experience

In addition to exploring the training effectiveness for training-oriented games, this study further investigates the game experience. Firstly, the reliability of the questionnaire was confirmed with a Cronbach’s alpha value of 0.95 for the training games group and 0.93 for the classic games group, demonstrating the trustworthiness of the questionnaire results. Subsequently, an ANOVA was conducted to compare participants' perceptions of training games and classic games, including three dimensions: flow, positive affect, and behavioral intention. The results showed no significant differences in these dimensions between the two groups, indicating that games designed for training purposes and classic games can both achieve a positive user experience. The results are presented in Table 2.

Table 2. Comparison of game experience between groups

<table>
<thead>
<tr>
<th></th>
<th>Training games (n=19)</th>
<th>Classic games (n=19)</th>
<th>F</th>
<th>etat²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow</td>
<td>3.64</td>
<td>3.16</td>
<td>1.96</td>
<td>2.25</td>
</tr>
<tr>
<td>Positive affect</td>
<td>3.72</td>
<td>3.49</td>
<td>0.61</td>
<td>0.53</td>
</tr>
<tr>
<td>Behavioral intention</td>
<td>3.22</td>
<td>3.32</td>
<td>0.10</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Homer et al., (2019) conducted an empirical study on digital training games and found that the game design has an impact on user preference and long-term usage intentions. In other empirical studies, Wells et al. (2021) compared the performance of EF training games with a control group and found that EF training-oriented designs have lower user affect and long-term usage intentions. The study also further emphasizes the impact of game elements, including challenges, on user experience. As the prior study has pointed out, enhancing EF requires continuous training, and users’ affective experience and long-term usage intentions should be given more attention. Therefore, this study aims to design EF training-oriented games that not only focus on training effectiveness but also emphasize the affective domain. The training process incorporates gaming elements such as level design, time limits, and scores, to gain challenges. Also, we use other game elements such as scenarios and real-time feedback in sound and animation to achieve a positive user experience in training-oriented games. This study aims to bridge the research gap and guide users toward continuing their training to achieve long-term goals in improving EF.

5. Conclusion

The executive function (EF) controls important cognitive abilities and can impact learning performance and daily life. Although prior studies developed digital training and training games to enhance EF, the affective domain is yet to be improved (Wells et al., 2021). Moreover, game scenarios of most EF training games are often far from real-life, which also influences the training process (Gray et al., 2019). Therefore, this study developed a training-oriented digital game with real-life scenarios to address this research gap. Through empirical research, this study answers the following research questions: (1) Does the training-oriented digital game enhance inhibition ability more effectively than classic games? ANCOVA results show that inhibition ability significantly improved through training with the training-oriented digital game compared to the classic game, demonstrating the effectiveness of this training-oriented digital game design. (2) Does the training-oriented digital game produce a more positive gaming experience than classic games? The result of analysis through questionnaires shows no significant differences in the flow, positive affect, and behavioral intention dimensions between the training-oriented game and the classic game. This indicates that both types of games can achieve the goal of a positive gaming experience and the training-oriented game can guide users to continue training to enhance EF. This provides insights into the design of EF training games, the game elements and situation should be emphasized.

The present study has some limitations and suggestions for future research. Due to the constraints of time, the sample size of this study was limited. Therefore, future studies should focus on larger sample sizes to improve the appropriateness of research inference. In addition, the training time was only 15 minutes. Although the design was based on previous research and was found to be effective, it is recommended to increase the training duration or frequency. Also, as the effectiveness of enhancing inhibition has been proven, the training aim of other EF sub-skills should be considered in future game development and studies. Finally, the training game was designed with scenarios close to real life, and the study mainly discussed its effectiveness and gaming experience. Research further comparing real-life and virtual situations is expected to establish the importance of situation setting.

References


Incorporating tangible rewards into gamification increases students’ identified regulation in fully online learning

Ya XIAO*, Khe Foon HEW*

*aFaculty of Education, The University of Hong Kong, Hong Kong SAR, China
*yaxiao@connect.hku.hk

Abstract: Intangible rewards (such as virtual points and virtual badges) have often been used in gamified learning contexts to motivate learners. However, such intangible rewards are not stimulating to all learners. Studies have reported that learners express their desire to redeem intangible rewards for some utilitarian resources or benefits. Considering that empirical evidence regarding how tangible rewards impact students’ motivation is still lacking, the present study applied a randomized controlled trial approach to explore the effects of tangible rewards on students’ identified regulation (a type of autonomous motivation) and external regulation (a type of controlled motivation). The study was conducted in a fully online gamified flipped class. Individual students were randomly assigned either to the tangible rewards group (EG = 28) or the intangible rewards group (CG = 29). Students in EG reported significantly higher identified regulation than those in CG while no significant difference was found in terms of external regulation. The results theoretically supported the standpoint that tangible rewards can help promote autonomous/self-determined motivation and gave practical guidance for educators who are interested in using tangible rewards in online gamified courses.

Keywords: Gamification, redeem, tangible rewards, motivation

1. Introduction

Gamification has been used to enhance greater student engagement, especially in the online learning environment (Zainuddin et al., 2022). Gamification rewards (Nicholson, 2015), usually take two forms — intangible and tangible. Intangible rewards refer to virtual badges/points and verbal praise that do not contain any utilitarian benefits to recipients (Meder et al., 2018). Tangible rewards refer to those that involve material goods (e.g., money) or some utilitarian benefits, (e.g., opportunities to participate in preferred activities) (Cameron et al., 2001; Kappen & Orji, 2017).

In most previous gamified learning studies, intangible rewards are more commonly used (Bai et al., 2021; Meder et al., 2018). Although non-utilitarian rewards can bring about enjoyable experiences in the gamified learning setting (Landers et al., 2015), not all learners favor these intangible factors in the long run. Instead, they tend to value the option of exchanging such rewards for more utilitarian or material resources (e.g., Huang & Hew, 2018).

2. Related Literature

2.1 Theoretical Controversy

The use of tangible rewards in education has been controversial and remains up for debate (Kappen & Orji, 2017). The controversy mainly revolves around how tangible rewards impact students’ autonomous motivations.

The self-determination theory makes a distinction between autonomous and controlled motivations based on their underlying regulatory processes and their associated
levels of self-determination. Autonomous motivation is a type of motivation in which individuals engage in a behavior because they find it inherently interesting, enjoyable, or meaningful. Controlled motivation, on the contrary, involves feeling pressured or coerced to engage in a behavior or activity due to external factors. Identified regulation, is a type of autonomous motivation that involves people identifying with the personal value and importance of the behavior for themselves and thus accepting it as their own. External regulation is a type of controlled motivation that involves people engaging in a behavior or activity solely to obtain a specific external outcome or reward (Ryan & Deci, 2000).

Specifically, some theorists criticize the detrimental effect of tangible rewards because tangible rewards tend to be perceived as controlling (i.e., being forced to complete designated tasks because of the reward). The perceived locus of causality changes from oneself to external rewards (Deci et al., 1975). They may begin to view the engagement in the designated tasks as a means to an end rather than as an inherently meaningful pursuit. Namely, they tend to feel that their efforts are only worthwhile if they are rewarded, rather than because they are improving their abilities. In this case, giving tangible rewards can lead to a decrease in identified regulation and an increase in external regulation.

However, other theorists argue that tangible rewards can be helpful if used appropriately. According to the social learning theory, when the rewards are tied to specific levels of performance (e.g., exceeding a certain score or surpassing other people's scores), an individual's perceived competence or self-efficacy will be enhanced (Bandura, 1986). The tangible rewards act as a social validation of one's competence, whereby a person's talents and abilities (internal factors) are perceived as the reasons for the rewards, rather than the rewards being perceived as the reasons for high-level performance (Cameron & Pierce, 2002). As a result, individuals' identification with the learning activity is strengthened instead of being harmed.

2.2 Previous Studies

Only a limited number of studies that have directly compared the use of tangible and intangible rewards in gamified educational contexts. For example, Ortega-Arranz et al. (2019) conducted a study that compared the behavioral engagement of students in a Massive Open Online Courses (MOOC) course among three groups: a "gamified group" (awarded with virtual badges only), a “tangible reward gamified group” (awarded with virtual badges redeemable for certain learning benefits), and a control group (without rewards). No significant difference was found between the two gamified groups although both of them outperformed the control group.

More recently, Bai et al. (2021) explored the effects of tangible rewards on students’ behavioral engagement and learning performance in a fully online university course, comparing students who were given virtual points to students who were given virtual points redeemable for high-quality learning materials. While tangible rewards encouraged students to create and reply to more posts, no significant difference was observed in their learning performance.

In general, there is still a lack of research that directly compares the effect of tangible and intangible rewards in fully online learning setting. In addition, the current empirical studies on gamification have not explored the effects of tangible rewards on students’ motivation.

3. Research Questions

The present study aims to fill the research gap by examining two research questions:

RQ1: What is the effect of tangible rewards on students’ identified regulation in online learning?
RQ2: What is the effect of tangible rewards on students’ external regulation in online learning?
4. Methodology

4.1 Context and Participants

The present study was conducted in the higher education context, in a fully online International Business course. The course comprised of eight sessions, each lasting 3.5 hours, and was aimed at preparing undergraduate students for a postgraduate entrance examination. Students who planned to take the examination and enter an international business master’s degree program voluntarily signed up for the course.

Altogether 57 participants with an average age of 21 were enrolled in this course. The participants were individually randomly assigned to either the control group (CG) or the experimental group (EG). The only difference between the two groups was that the EG students could redeem exclusive learning materials with their virtual points while the CG students could not. All the redeemable rewards were later given to the CG students after the experiment to avoid ethical issues.

4.2 Gamified Flipped Class Design and Redemption Scheme

In a typical flipped classroom, students engage in (1) pre-class computer-based learning (such as video lectures) and (2) in-class interactive learning. Students are expected to familiarize themselves with pre-class learning materials and focus on higher-level cognitive activities, such as peer learning, and problem-solving (Bishop & Verleger, 2013; Sointu et al., 2023).

To increase student engagement in the pre-and in-class learning activities, points, badges, and leaderboards (known as the PBL triad) were used for both CG and EG. The PBL triad is one of the most commonly used gamification elements (Leitão et al., 2022). Please see Table 1 for a detailed gamification design.

<table>
<thead>
<tr>
<th>Points</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5 points for downloading pre-class learning materials (only the first download will be counted)</td>
</tr>
<tr>
<td>10</td>
<td>10 points for answering each pre-class MCQ question correctly (5 MCQ questions in total)</td>
</tr>
<tr>
<td>5</td>
<td>5 points for completing all pre-class learning activities, including downloading the pre-class learning activities and fulfilling pre-class tests</td>
</tr>
<tr>
<td>10</td>
<td>10 points for answering each in-class MCQ question correctly (10 MCQ questions in total)</td>
</tr>
<tr>
<td>10</td>
<td>10 points for voluntarily answering or posing questions in class (only the first three attempts will be counted)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Badges</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Prepared</td>
<td>This badge rewards students whose accuracy rates in the pre-class quizzes are greater than or equal to 80%</td>
</tr>
<tr>
<td>Task Completion</td>
<td>This badge rewards students who complete all pre-class learning activities, including downloading the pre-class learning activities and fulfilling pre-class tests</td>
</tr>
<tr>
<td>Task Master</td>
<td>This badge rewards students who complete the ten items in the in-class quizzes</td>
</tr>
</tbody>
</table>
**Leaderboards**

**Description**

Leaderboard by session
This leaderboard announces students' rankings and accumulated points by each weekly session.

Overall Leaderboard
This leaderboard announces students' overall rankings and accumulated points throughout the whole course.

**Tangible rewards**

**Description**

Exclusive learning materials
(SE only) The rewards could be redeemed weekly for students who accumulated 160 virtual points or above that week.

In this study, the course was conducted through a synchronous videoconferencing tool. All the pre- and in-class learning activities were conducted using a self-designed gamified learning platform.

**4.3 Data Collection and Analysis**

Students identified regulation and external regulation was measured by scales adapted from the Situational Motivation Scale (SIMS, Guay et al, 2000). The Shapiro–Wilk statistics showed a non-normal distribution of student responses in the CG ($W = 0.924, p = .038$) and the EG ($W = .846, p = .001$) in terms of identified regulation. Similarly, in terms of external regulation, student responses in the CG ($W = .917, p = .025$) and EG ($W = .915, p = .026$) were also non-normally distributed. In this case, students' responses to these two scales were examined using the Mann–Whitney U test.

**5. Results**

**5.1 Effects on Students' Identified Regulation**

The questionnaire results indicated good reliability (Cronbach's alpha = 0.947). As shown in Table 3, the students in the EG ($Mdn = 6.167, SD = 1.044$) reported a significantly higher level of identified regulation than those in the CG ($Mdn = 4.667, SD = 1.107; U = 630.000, p < .05, r = 0.478$).

**5.2 Effects on Students' External Regulation**

The questionnaire results indicated relatively acceptable reliability (Cronbach's alpha = 0.617) (Hajjar, 2018; Wim et al., 2008). The students in the EG ($Mdn = 3.333, SD = 1.373$) reported no significant difference in terms of external regulation with those in the CG ($Mdn = 4.000, SD = 1.119; U = 330.000, p = .222$) (see Table 3).

<table>
<thead>
<tr>
<th>Measures</th>
<th>Comparison group</th>
<th>$Mdn$ (SD)</th>
<th>$p$-value (effect size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified regulation</td>
<td>EG</td>
<td>6.167 (1.044)</td>
<td>$p = .000^*, r = 0.478$</td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>4.667 (1.107)</td>
<td></td>
</tr>
<tr>
<td>External regulation</td>
<td>EG</td>
<td>3.333 (1.373)</td>
<td>$p = .222$</td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>4.000 (1.119)</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* * significant using $p < .05$
6. Discussions and Conclusions

This study applied a randomized controlled trial approach to examine the effects of tangible rewards redeemed through intangible rewards on students’ identified regulation and external regulation in a fully online gamified flipped class. The results showed that the use of tangible rewards significantly increased students’ identified regulation while indicating no significant influence on students’ external regulation.

6.1 Implications

The findings shed more light on the theoretical controversy over the effects of tangible rewards. Specifically, the increase in identified regulation indicates that tangible rewards facilitate the internalization of motivation — the process by which an individual adopts and integrates an external regulation or value into their sense of self (Deci & Ryan, 2015). Both the reward scheme and the reward type are crucial to the effectiveness of tangible rewards in promoting internalization.

In this study, the redemption of tangible rewards is associated with a performance standard (i.e., earning over 160 virtual points, which is roughly 80% of the full virtual points for a week). This criterion provides valuable information that can positively reinforce students’ competence (Bandura, 1986), thereby mitigating the potential negative effects of the reward being a controlling factor (Ryan & Deci, 2000). Additionally, according to the extended attribution theory, tangible rewards tied to performance levels enhance perceived competence and encourage individuals to attribute their success to their own abilities (Lepper et al., 1996).

The type of tangible rewards used in this study is compatible with learning purposes, i.e., exclusive learning materials that may boost students’ scores in the postgraduate entrance examination. In other words, the redeemed tangible rewards are personally meaningful/relevant to students as they are closely related to their valued future goals. Such personal relevance is crucial to identified regulation (Vansteenkiste et al., 2018). That explains why in the redeeming procedure, students tend to identify learning activities with personal meaning, endorsing that learning is good for themselves.

To summarize, theoretically, this study provides evidence supporting the theoretical standpoint that tangible rewards can be helpful in promoting autonomous/self-determined motivation if used appropriately. Practically, this study provides two important implications for applying tangible rewards in the gamified learning context, (1) linking tangible rewards to a standard of performance; (2) linking tangible rewards to students’ valued learning goals.

6.2 Limitations

Three primary limitations need to be acknowledged in this study. Firstly, the research focused solely on a fully online class that utilized synchronous videoconferencing technology. Therefore, the outcomes of this investigation may not be applicable to other learning contexts, such as asynchronous learning in MOOCs or face-to-face learning. Secondly, the tangible rewards implemented in this study may only be effective for students who are preparing for the entrance examination. The use of this kind of reward may not be generalizable to other situations, highlighting the importance of providing rewards that match the specific needs of the target population. To address this issue, future research could include a pre-survey to gather information about students’ requirements and preferences. Thirdly, the relatively small ample size may limit the generalizability of the findings.

Acknowledgements

The research was supported by a grant from the Research Grants Council of Hong Kong (Project Reference No: 17616020).
References


Effects of different embodied scaffoldings on students' spatial abilities in digital game-based learning

Junyi ZHOU\textsuperscript{a}, Jialing ZENG\textsuperscript{b}, John B. BLACK\textsuperscript{b} & Junjie SHANG\textsuperscript{a}\*  
	extsuperscript{a}Lab of Learning Sciences, Graduate School of Education, Peking University, Beijing, China  
	extsuperscript{b}Teachers College, Columbia University, New York, America  
\*jishang@pku.edu.cn

Abstract: Adding embodied scaffolding to the teaching process can be effective in improving students' spatial abilities. However, few studies have been conducted to explore the effects of different embodied scaffoldings on students' spatial abilities. Therefore, the purpose of this study was to investigate the effects of different embodied scaffoldings on students' spatial learning. Three types of embodied scaffoldings were designed: controllable animation scaffolding (CA), instructional gesture plus animation scaffolding (GA), and physical object scaffolding (PO). Based on this, we conducted a quasi-experimental study in which 197 elementary school students were randomly divided into three groups to learn geometry in a game-based learning context and measured their knowledge learning, spatial ability, and flow experience. The results showed that although there were no significant differences in spatial ability among the three groups, students in the PO group performed significantly better in knowledge transfer than the other two groups, while students in the PO group also had the lowest level of flow experience. In addition, we investigated the influence of students’ prior proficiency on the effects of different embodied scaffoldings. The results showed that students in the low proficiency group performed better in the PO condition and the CA condition than those in the high proficiency group.

Keywords: Spatial ability; Game-based learning; Embodied cognition; Scaffolding

1. Introduction

Spatial ability can be defined as a human’s capacity to understand, reason over, recall and manipulate the spatial relations among objects or in space, consisting of three factors: mental rotation, spatial perception, and spatial visualization (Linn & Petersen, 1985). As one of the basic human cognitive abilities, spatial ability is important for people to recognize their own environment and solve spatial problems (Duffy, Sorby, & Bowe, 2020; Gardner, 1983). Spatial ability levels not only directly affect learners’ understanding, representation, and solution of science, technology, engineering, and mathematics (STEM) problems, (He, et al., 2021; Hodgkiss, et al., 2018) but also predict learners’ career achievement and career choices in STEM fields (Yoon & Mann, 2017).

In the current pedagogical community, game-based learning is widely adopted by many educators. Numerous studies have shown that educational game is effective in enhancing learners' learning performance (Cerra, et al., 2022), self-efficacy (Kuznetcova, et al., 2023), and motivation (Fadda, et al., 2022). In addition, educational games can provide students with visual representations of objects in a three-dimensional form, and students can master spatial relations in the process of interacting with games, which also has a positive impact on spatial ability (Uttal, 2000). Therefore, some researchers have tried to integrate digital game-based learning into the teaching and learning process of spatial ability and have confirmed the effectiveness of educational games in improving students' spatial ability (Hou, et al., 2021; Chai, et al., 2019; Lin & Chen, 2016).
However, current educational games for developing students’ spatial ability pay less attention to the essential features of spatial cognition and neglect the importance of embodied cognition for spatial ability. Evidence from several studies suggests that there is a link between the body, action, and spatial cognition and that physical movement plays an active role in spatial cognition (Lozano, Hard, & Tversky, 2007; Morsella & Krauss, 2004). By offloading mental processing to physical actions, it may help to improve students’ understanding of spatial relationships. Some studies have been conducted to design teaching activities for spatial abilities based on the perspective of embodied cognition, and the results have shown that students who use embodied scaffolding achieve better learning outcomes (Kwon, et al., 2023; Rabattu, et al., 2023; Burte, et al., 2017). Therefore, given the value of scaffolding for student learning, educators may consider introducing embodied scaffolding to support students’ spatial development.

Currently, few existing studies have explored the best practice of designing embodied scaffolding to enhance spatial ability in digital game-based learning. Therefore, this study aims to investigate and compare the effects of three different types of embodied scaffolding on students’ spatial learning, with further consideration of the influencing factor, students’ initial proficiency.

2. Literature Review

2.1 Embodied scaffolding in digital game-based spatial ability learning

The theory of embodied cognition states that the process of knowledge construction is inseparable from the physical interaction of the learner with the learning environment (Ioannou & Ioannou, 2020). By expanding learning from visual cognitive activities to physical movement, the researchers believe that this allows students to engage in multimodal interactions that help students process learning content at a deeper level (Anastopoulou et al., 2011) and understand abstract scientific concepts. In addition, several studies have shown that embodied learning enhances students’ spatial thinking skills (Burte et al., 2017) and that the embodied element is positively associated with mathematical proof performance, insight, and intuition (Nathan et al., 2021).

Based on this, a growing number of researchers have attempted to apply embodied cognition theories to instructional design. Some studies encourage students to increase physical movement as an aid to cognitive processes (Rollinde, et al., 2021), while others incorporate the idea of embodied cognition into scaffolding to support their learning (Rahimi, et al., 2022). This scaffolding, which is provided during instruction to help students offload cognition onto the body or external objects, is referred to as embodied scaffolding. Embodied scaffolding transforms sensory experience into cognition and reduces cognitive load through physical movement (Zhang, et al., 2022) to help learners integrate abstract concepts with the learning environment and achieve better learning outcomes (Rahimi, et al., 2022).

Embodied scaffolding can be represented in many forms, and there are also differences in the motor nerves that can be mobilized by various embodied scaffoldings. Controllable animation is one of the most common embodied scaffoldings. Controllable animation scaffolding is a less embodied form of learning support in which learners can use finger tracking to control the animation with tools such as virtual sliders to achieve embodied cognition. For example, Johnson-Glenberg et al. (2016) designed animations that can control the speed of a round ball through a virtual slider to assist students in learning physics. In addition, some researchers have pointed out that gestures can act as a cross-modal prime (Hostetter & Alibali, 2008), providing learners with additional memory codes, thereby strengthening memory representation, increasing retrieval cues (Johnson-Glenberg & Megowan-Romanowicz, 2017) and reducing the cognitive load of learners (Goldin-Meadow, 2014). Therefore, some research has also attempted to combine instructional gestures with animation to mobilise more motor nerves while promoting students’ embodied cognition. For example, in Merkouris et al. (2019)’s study, students remotely manipulated programmed robots with the help of instructional gesture scaffolding and animation scaffolding. Besides,
physical object scaffolding is a highly embodied learning support in which learners can gain a
deeper tactile experience by interacting with teaching aids, joysticks, force feedback devices,
etc., and offload their cognition to the physical object. For example, Zohar et al. (2021) used
a haptic device that applied force feedback to help students learn about chemical bonding.

In summary, researchers have designed rich embodied scaffolding in different
disciplines to support students' embodied learning. However, few studies have attempted to
investigate embodied scaffolding for spatial ability learning, leaving a certain gap in this area.
In addition, most of the existing research has been situated within traditional teaching
approaches, and no research has attempted to explore the effect of different embodied
scaffolding in the context of game-based learning.

2.2 Effect of initial proficiency on embodied learning

When designing embodied interventions, it is necessary to fully consider the relevant factors
of the learners themselves, such as prior knowledge or ability (Conley et al., 2020; Post et al.,
2013). Chi and Glaser (1985) noted that the level of expertise of the learner is a key factor in
determining what information is relevant to the learner and what information to focus on. As
learners increase in proficiency, their embodied learning styles may change somewhat.
Novices, for example, often use their fingers when engaging in abstract number tasks, but as
students’ expertise increases, mental representations become more abstract and simplified
(Pouw et al., 2014), so experts prefer to solve problems in an intangible way. The Expert
Reversal Effect also points out that instructional guidance that is essential for novices may
negatively affect more experienced learners, meaning that learners with less domain
knowledge are more likely to benefit from an environment with a large number of sources of
information and instructional support, while the opposite is true for learners with higher domain
knowledge (Kalyuga et al., 2003).

There is empirical evidence of the effects of initial proficiency on embodied learning. Pouw et al. (2016) found that children with lower math skills benefited more from using augmenting instructional animations with a body analogy (BA) compared to children with higher math skills. Similar evidence was found in statistical disciplines (Conley et al., 2020) and computational thinking (Merkouris & Chorianopoulos, 2019) that participants with low prior knowledge may benefit most from higher levels of embodied experience, while participants with high prior knowledge prefer disembodied and abstract learning content. However, there are studies that have come to the opposite conclusion. Swart et al. (2017) found that students with lower initial proficiency benefited more from playing the deictic gesture (pointing) version of the game, while students with higher initial proficiency benefited more from the iconic gesture (metaphorical, enactive, symbolic) version of the game. Therefore, students’ proficiency also need to be considered when designing embodied scaffolding.

2.3 The Present Study

The above studies have discussed the effects of embodied scaffolding on digital game-based
spatial ability learning. However, few studies have focused on the effects of different embodied
scaffolding on improving spatial ability. In addition, little or no research has considered the
influence of learners’ prior proficiency. For this reason, a quasi-experimental study was
conducted to assess the effects of different embodied scaffolding on students' spatial ability
learning and to examine the effects of learners’ prior proficiency on embodied learning. Three
common types of embodied scaffolding were selected for this study: controllable animation
(CA), instructional gesture plus animation (GA), and physical object (PO). The following two
research questions were explored:

1. What is the effect of different embodied scaffolding on students' learning of spatial
   ability?
2. Do students with various proficiency perform differently with each embodied
   scaffolding support?
3. Method

3.1 Participants

The participants were 197 fourth graders (100 females, $M_{\text{age}} = 10.09$, $SD_{\text{age}} = 0.86$) from three classes in one public primary school in Sichuan, China. All students had not been exposed to the learning content and the ANOVA analysis of three classes’ math final exam grades of last academic semester showed no significant difference ($F(2, 191) = 0.437, \rho = .647, \eta^2_p = 0.005$). Previous studies have confirmed the correlation between spatial ability and mathematical ability (Geer et al., 2019), so we concluded that the initial spatial ability of the students in the three classes was similar. The classes were randomly assigned to use three embodied scaffolding: CA condition (32 females, 33 males) to learn using the controllable virtual three-dimensional simulation, GA condition (35 females, 31 males) to learn using the simulation plus instructional gestures, and PO condition (33 females, 33 males) to learn using physical objects. Each class was taught by the same two instructors with extensive teaching experience.

3.2 Procedure

The experiment was conducted as an 11-day experiment. All students took a three-day lesson (40 min per day) about mental folding from day 8. A ten-minute basic concept instruction was given in the first class and then students used the digital game environment on tablet to learn by themselves, with instructors providing learning supports. Thirty-two-minute pre-, post-tests were administered to students on the first and the last day of the experiment, respectively. Furthermore, students completed the flow questionnaire after the last lesson.

3.3 Materials

The digital educational game was called Cube Elimination, which focused on spatial visualization ability training and math geometric content learning of the cube nets fold and unfold. The Cube Elimination game was designed to help students form spatial representations, build an understanding of the cube and net concepts, and construct the association between three- and two-dimensional objects. Specifically, students were required to complete three main challenges presented in the game: 1) Given a net with the same pattern on each side, and find the opposite sides of the shaped cube; 2) Given a net marked with bottom/top surface, and find the front and top/bottom surface of the formed cube; 3) Given a cube, find the corresponding net from a figure. None of the students had prior experience with Cube Elimination.

The embodied learning resources varying in scaffolding were as follows. In the CA condition, the game provided the student with a three-dimensional cube that can be rotated and viewed, as well as a dynamic animation of the transition between the cube and the net. The student can control the speed, progress, and so on of the animation (see Figure 1a). In the GA condition, students were not only exposed to the three-dimensional cubes and dynamic animations in the game, but were also guided to use gestures that represented the transition process (see Figure 1b). In the PO condition, the student would get the physical magnetic square pieces (see Figure 1c).
3.4 Measures

The pre-tests and post-tests included questions assessing two categories of variables: spatial ability (mental folding and mental rotation) and knowledge learning (retention and transfer). Students' knowledge learning status can reflect their understanding of objective knowledge in the process of spatial ability cultivation, so it was also measured in this study. Moreover, the flow questionnaire contained items collecting data on students’ learning experiences.

3.4.1 Spatial ability

The spatial ability test included mental folding test, which was the ability that students were trained directly in the game, and mental rotation test, which was the transfer ability that students were not trained directly. For the mental folding test, the Paper Folding Test from the Kit of Factor Referenced Cognitive Tests (Ekstrom et al., 1976) was used, containing 20 items. In this test, students were given a square paper that had been folded and punctured in a series, as well as five figures showing the position of the holes when the paper was fully unfolded. They were asked to find the correct number from among the five figures. Each correct answer was worth one point, while each incorrect or no answer was worth 1/5 of a point, on a scale from -4 to 20.

The Mental Rotation Test was administered using the Vandenberx & Kuse Mental Rotation Test (Peters et al., 1995) and contained 24 items. Each test contained a target figure as well as four stimulus figures, and students were asked to select the rotated version of the stimulus figure that matched the target figure. If the correct stimulus figure was found, one point was given, on a scale from 0 to 24.

3.4.2 Knowledge learning

The knowledge learning test contained questions from the retention test and the transfer test, with each including ten questions, a total of 20 points. The knowledge test lasted 20 minutes. The retention test measured students' understanding of cubes and nets. For example, given a numbered network, students were asked to find the opposite side of the number two. The transfer test evaluated students' application of knowledge in new situations, such as asking students to find shapes that cannot be folded into cubes. On both the retention test and the transfer test, each correct answer received one point, while each incorrect or no answer received zero point. Pre-test and post-test questions were different, but the content and difficulty were similar. In addition, the knowledge test was tested by two experienced math teachers and a math teaching expert, with good expert validity.
3.4.3 Flow experience

The Short Flow State Scale-2 (SFSS-2) developed by Jackson, Ecklund and Martin (2008) was used as a flow experience questionnaire to measure the intensity of flow experience in students in game-based embodied learning. The scale consisted of nine items corresponding to the nine-dimensional conceptualization of convection by Csikszentmihalyi (1990). SFSS-2 was rated on a six-point Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree), with higher scores indicating stronger feelings of flow. The reliability of SFSS-2 was $\alpha=0.77$ (Jackson et al., 2008).

3.5 Data analysis

The data was analyzed in SPSS v.28 (IBM Corp, 2021). To answer RQ1, analyses of variance (ANOVA) were conducted to assess differences between conditions in post-test scores. As the result of an ANOVA showed that there was a significant difference between conditions in the pre-test for mental folding ability (F (2, 194) = 3.646, $p = 0.028^*$, $\eta_p^2 = 0.036$), a univariate analysis of covariance (ANCOVA) was employed to account for mental folding pre-test scores. All variables met the assumption of homogeneity of variance given the results of Levene’s tests, except the flow experience. Thus, Welch’s ANOVA was used for the analysis of the flow variable. Table 1 presents the descriptive statistics for all variables.

To answer RQ2, a median split was adopted to assign students in each condition to distinct levels of prior proficiency (Bobek & Tversky, 2016). Based on students’ pre folding test scores, they were divided into low- and high-proficiency groups. Table 2 sets out the descriptive statistics for each group in three conditions on all variables. There were significant differences between the low and high prior proficiency groups in each condition (CA: F (1, 56) = 164, $p < .001^{** *} , \eta_p^2 = 0.745$; GA: F (1, 51) = 88.0, $p < .001^{***}, \eta_p^2 = 0.633$; PO: F (1, 58) = 185, $p < .001^{***}, \eta_p^2 = 0.762$). To compare each group’s improvement from the pre-test to the post-test, we calculated the learning gain by subtracting the pre-test score from the post-test score. ANOVAs were employed to examine differences in learning gains between groups under each condition. All variables met the assumption of homogeneity of variance or had equal group sizes.

4. Results

4.1 Embodied scaffolding

With respect to the effect of embodied scaffolding on knowledge learning after classes, the three conditions did not significantly differ in their retention scores, F (2, 194) = 0.673, $p = 0.511, \eta_p^2 = 0.007$, whereas the analyses revealed a significant effect of condition on transfer scores, F (2, 194) = 3.603, $p = 0.029^*$, $\eta_p^2 = 0.036$. The follow-up LSD comparisons showed that students in the PO condition performed better in the transfer test than students in the CA condition, MD = 0.88, SD = 0.374, $p = 0.19^*$, and students in the GA condition, MD = 0.85, SD = 0.372, $p = 0.024^*$, respectively.

Regarding the effect of embodied scaffolding on spatial abilities after classes, the analyses did not find a significant effect of condition on mental folding ability, F (2, 193) = 1.578, $p = 0.209, \eta_p^2 = 0.016$, nor on mental rotation ability, F (2, 194) = 2.967, $p = 0.054, \eta_p^2 = 0.030$.

Regarding the effect of embodied scaffolding on students’ flow experience during learning, the analyses revealed a significant effect of condition on flow scores, F (2, 127.42) = 8.807, $p < .001^{***}, \omega_p^2 = 0.061$. The results of follow-up analyses indicated that students in the PO condition had a lower feeling of flow than those in the CA condition, MD = 0.53, SD = 0.137, $p < .001^{***}$, and students in the GA condition, MD = 0.31, SD = 0.137, $p = 0.027^*$.

Table 1. Descriptive statistics for pre-, and post-tests
<table>
<thead>
<tr>
<th>Variable</th>
<th>CA Condition (N=65)</th>
<th>GA Condition (N=66)</th>
<th>PO Condition (N=66)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td></td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
</tr>
<tr>
<td>Retention</td>
<td>3.46(2.13)</td>
<td>6.88(1.97)</td>
<td>3.70(2.00)</td>
</tr>
<tr>
<td>Transfer</td>
<td>4.65(1.92)</td>
<td>5.65(2.19)</td>
<td>4.55(2.06)</td>
</tr>
<tr>
<td>Folding</td>
<td>4.97(4.46)</td>
<td>7.85(4.57)</td>
<td>4.64(3.81)</td>
</tr>
<tr>
<td>Rotation</td>
<td>6.28(3.15)</td>
<td>7.63(3.53)</td>
<td>7.17(2.88)</td>
</tr>
<tr>
<td>Flow</td>
<td>4.76(0.65)</td>
<td>4.54(0.90)</td>
<td>4.23(0.78)</td>
</tr>
</tbody>
</table>

4.2 Prior proficiency

With respect to the pre-post test learning gain, the analyses of knowledge retention data showed that there was no significantly greater gain between the low proficiency group and the high proficiency group (CA: F (1, 56) = 0.0118, p = 0.914, ηp² = 0.000; GA: F (1, 51) = 1.16, p = 0.286, ηp² = 0.022; PO: F (1, 58) = 0.00, p = 0.999, ηp² = 0.000). Regarding the knowledge transfer, the results still failed to reach statistical significance (CA: F (1, 56) = 0.118, p = 0.732, ηp² = 0.002; GA: F (1, 51) = 0.0798, p = 0.779, ηp² = 0.002; PO: F (1, 58) = 0.407, p = 0.526, ηp² = 0.007).

With regard to the mental folding ability, the analyses showed that the low proficiency group achieved a significantly greater gain than the high proficiency group in the CA condition, F (1, 56) = 11.5, p = .001**, ηp² = 0.171, as well as in the PO condition, F (1, 58) = 9.72, p = 0.003**, ηp² = 0.144. In the GA condition, although the low proficiency group had a numerically greater gain than the high proficiency group, the results failed to find a statistical significance, F (1, 51) = 1.32, p = 0.257, ηp² = 0.025. Regarding the mental rotation ability, results showed that there was no significant greater gain between the low proficiency group and the high proficiency group (CA: F (1, 56) = 1.46, p = 0.232, ηp² = 0.025; GA: F (1, 51) = 1.12, p = 0.295, ηp² = 0.021; PO: F (1, 58) = 2.34, p = 0.131, ηp² = 0.039).

5. Discussion

5.1 Effects of embodied scaffolding on spatial ability learning

This study assessed students’ spatial ability, knowledge learning, and flow experience after game-based instruction in three conditions. In terms of spatial ability learning, the main results showed that while students’ spatial ability improved significantly in all three embodied scaffolding conditions, there were no significant differences in scores on mental folding and mental rotation in the three conditions. This may verify the importance of incorporating embodied scaffolding in teaching spatial ability (Kwon, et al., 2023; Rabattu, et al., 2023; Burte, et al., 2017), regardless of the form of scaffolding.
For geometry knowledge learning, the main results showed that there was no significant difference in knowledge retention scores between the CA condition, the GA condition, and the PO condition, while students in the PO condition performed significantly better in the transfer assessment of geometry knowledge. The results showed that all three types of embodied scaffolding were able to support students’ learning of geometry while also demonstrating the advantages of object embodied scaffolding in terms of knowledge transfer. This is consistent with previous studies suggesting that engaging in higher levels of embodied learning activates sensorimotor codes that enhance memory traces and help learners learn content faster and in greater depth (Johnson-Glenberg et al., 2020). As physical object scaffolding is a relatively high-embodiment scaffolding, students would gain more diverse memory cues when using it, which may help them to extract knowledge more quickly on transfer tests.

In terms of students’ flow experience, the main results showed that the scores of flow experience were significantly lower in the PO group than in the other two groups. The reason may be that students in the PO group had to switch frequently between electronic devices and magnetic square pieces during the learning process, and this shift in attention may have affected the students’ flow experience.

5.2 Performance of students with different proficiency in embodied scaffolding

To investigate whether students with different proficiency would perform differently with each embodied scaffolding support, this study divided students into high and low proficiency groups based on the pre folding test scores. Analysis of the learning performance of the two groups revealed that there was no significant difference in the learning gain in terms of knowledge retention, knowledge transfer and mental rotation ability between the two groups, regardless of the embodied scaffolding used. As for the mental folding ability, there was a significant difference in the learning gain between students in the high proficiency group and students in the low proficiency group in both the CA condition and the PO condition, with both showing greater progress for students in the low proficiency group. The reason may be that students in the CA group and the PO group were able to get clear feedback from the scaffolding during the learning process, while students in the GA group were unable to determine whether their movements were consistent with the real folding process. On the other hand, imagining the process of folding the cube out of thin air also tests students’ spatial thinking skills, which may not be applicable to students with low initial spatial ability.

6. Conclusion

Our study investigated the effects of different embodied scaffoldings on students' learning of spatial ability. The results showed that students in the CA group, the GA group, and the PO group were able to learn geometry and enhance their spatial abilities, indicating the effectiveness of incorporating embodied scaffolding in enhancing spatial abilities. For the transfer assessment of geometric knowledge, the PO group scored significantly higher than the other two groups, although this group also scored significantly lower than the other two groups for flow experience. This suggested that we should not only explore the cognitive value of object scaffolding, but also pay attention to its negative impact on flow experience. In addition, we further analyzed the possible effects of students’ proficiency in different embodied scaffolding conditions and showed that students with lower proficiency were more suitable for CA scaffolding and PO scaffolding with clearer feedback. We hope that our work will help educators to better initiate the teaching of spatial ability and apply appropriate embodied scaffoldings to enhance students’ spatial ability.

Acknowledgements
We would like to acknowledge Ruonan Hu, Zhiyong Li, Yu Li, Xiaqi Wei, Yuxi Wei, Hua Tian, Songjiang Chen, Yongguo Wang, Yanjun Cao, Jing Feng, Yingyong Jin, Jing Zhao, Hailing Tang, Jing Li for their help with experiment conducting and data collection.

Foundation: Key Project of The Beijing 14th Five Year Plan in 2021 for Education Science, “Research on the Cultivation of Learning Ability based on Learning Sciences and Game-based Learning”(No.CEFA21067)

References


Exploring the Impact of Designing a Robot as a Pet with Interdependence Theory on Long-Term Relationships and Learning Performance

Vando Gusti Al HAKIM\textsuperscript{ab}, Su-Hang YANG\textsuperscript{c}, Jen-Hang WANG\textsuperscript{d}, Yu-Chen CHANG\textsuperscript{a}, Hung-Hsuan LIN\textsuperscript{a} & Gwo-Dong CHEN\textsuperscript{a}

\textsuperscript{a}Department of Computer Science and Information Engineering, National Central University
\textsuperscript{b}Department of Electrical Engineering Education, Yogyakarta State University, Indonesia
\textsuperscript{c}Department of Hospitality Management, Chien Hsin University of Science and Technology
\textsuperscript{d}Research Center for Science and Technology for Learning, National Central University
\textsuperscript{acd}Taiwan
\textsuperscript{*}harry@cl.ncu.edu.tw

Abstract: Educational robots have shown promise in enhancing learning performance; but, many of these robots function solely as companions or tutors and rely on novelty to attract attention, making it challenging to maintain long-term relationships. Although pet robots have been utilized to create relationships with users, their potential in education remains underexplored. This study addresses this issue by designing a robot as a pet, applying the interdependence theory to establish lasting relationships and improve learning performance. The interdependence theory suggests that relationships between individuals can be strengthened through mutual dependence and the fulfillment of each other's needs. In this study, learners were prompted to engage in continuous care for their pet robots, receiving emotional reinforcement from them seamlessly across both virtual and tangible forms enabled by the ChatGPT API. To culminate their learning process, learners presented their ultimate learning outcomes alongside their pet robots, enacting situational dramas for their classmates. To evaluate the effectiveness of this approach, a quasi-experiment was conducted with 100 undergraduate learners enrolled in a Japanese for Hospitality and Tourism course in Taiwan. Results indicate that robots designed as pets, leveraging interdependence theory, significantly yield positive effects on learners, encompassing improved learning outcomes, extended interaction rates, and heightened satisfaction with their study journey compared to conventional robots. The study concludes by discussing research limitations and offering suggestions for further enhancements to this approach.

Keywords: Educational Robots, Pet Robots, Long-Term Relationships, Interdependence Theory, Situational Learning, Human-Robot Interaction

1. Introduction

In recent years, the use of robots to support learning has gained significant attention among scholars (Belpaemae et al., 2018; Mubin et al., 2013). Previous studies have demonstrated the potential of robots to engage learners and stimulate their interest in interacting with learning materials (Al Hakim et al., 2021; 2022a; Lin et al., 2022). However, since learning often requires several weeks or months to complete, it is essential for robots to develop enduring relationships with learners to maintain their engagement for an extended period.

Establishing long-term relationships with robots has been identified as a significant factor in the realm of robotics (Westlund et al., 2018; Leite et al., 2013). This is because learning and behavioral changes often require continuous engagement and repetition, rather than just a single interaction. However, as the frequency of interactions increases, the initial novelty effect may diminish, potentially affecting the efficiency of learning.
One promising approach to cultivating long-term relationships is drawing inspiration from the bond between pets and their owners (Díaz-Boladeras, 2022). Research has shown that treating robots as pets can elicit positive emotional engagement and attachment (Westlund et al., 2018). The relationship between pet owners and their pets is well-documented, with pets providing emotional support and interacting with them fulfilling social needs, promoting happiness, and fostering healthier personality traits (McConnell et al., 2011). A well-known illustration is Tamagotchi, a virtual pet that sends reminders to users for their care (Lawton, 2017). Despite its simple design, users often interact with Tamagotchi extensively, creating a connection where they care for the pet and receive emotional support. In education, interdependence was studied by Johnson & Johnson (2009), highlighting its importance for successful ongoing collaboration.

Building upon the groundwork laid by Al Hakim et al. (2023) regarding pet-like robot models, our study bridges gaps by investigating their impact on education. Through a novel learning design rooted in interdependency, seamless integration of technology for ubiquitous learning, and a comparative evaluation, we aim to enhance educational robot design. Specifically, we explore learners’ potential to establish lasting relationships with pet-like educational robots. By imbuing robots with needs like feeding, we stimulate interaction rates. Furthermore, the robot’s presence may provide emotional support, enhancing satisfaction with the experience and covered material. This study is guided by the research questions:

1. Does a pet robot established with an interdependence mechanism enhance learning outcomes more effectively than a conventional robot?
2. Does a pet robot established with an interdependence mechanism increase interaction rates more effectively than a conventional robot?
3. How does the use of a pet robot with an interdependence mechanism, in comparison to a conventional robot, impact learners’ acceptance of learning?

2. Related Works

2.1 Robots in Education

Robots have emerged as a prominent area of research in education, witnessing significant growth and diverse applications. They have been explored as tutors or peer learners, yielding positive outcomes in cognitive and emotional outcomes comparable to human tutoring (Belpaeme et al., 2018). They can also serve various functions and be applied in different areas of education, including language, science or technology education (Mubin et al., 2013). Despite the numerous studies on educational robots, there is still a lack of study concerning the establishment of long-term relationships with these devices. To address this, it is crucial for robots to have specific roles and interaction modes that can sustain user engagement. Previous research has mainly focused on the role of educational robots as a tutor or friends with personal conversation (Ligthart et al., 2023), or storyteller companions (Calvo-Barajas & Castellano, 2022), which relies solely on social and affective behaviors.

2.2 Pet Robots and Long-Term Relationships

The introduction of Bandai’s "Tamagotchi" in 1996 marked a significant milestone in the acceptance of virtual pets, owing to their portability and sociability (Lawton, 2017). As a result, many studies related to artificial pet have become popular topics of study. Bylieva et al. (2019) conducted a literature review that traced the development of artificial pets from the 1990s to the 21st century. They identified social interaction as a crucial factor in sustaining relationships with artificial pets. However, the exact definition of a "long-term relationship" remains a topic of debate within the domain of human-robot interaction (Díaz-Boladeras, 2022). While there is no consensus on the specific duration or number of interactions required, a long-term relationship is generally believed to emerge when the user becomes accustomed to the robot and the novelty effect no longer influences their perception of it (Leite et al., 2013). Díaz-
Boladeras (2022) suggests measuring the long-term relationship based on changes in behavior, such as the rates of use, rather than relying solely on the time span. Furthermore, the relationship mechanism between a user and a robot is a vital aspect to consider. Frude and Jandrić (2015) introduced the concept of "animism," which suggests that humans tend to attribute life and consciousness to non-living objects, thereby fostering emotional bonds. This concept highlights the potential for users to form attachments with robots. Similarly, Heljakka et al. (2021) discovered that the relationship between users and robot dogs can deepen over time as they spend more time together. However, the precise mechanisms underlying the establishment and maintenance of such relationships have yet to be fully explored in these studies.

2.3 Interdependency Theory

The concept of interdependence was first introduced by John Thibaut and Harold Kelley in 1959, encapsulating elements such as dependence, power, norms, rules, cooperation, and coordination. It explains the extent to which two or more individuals rely on each other to achieve their goals within a specific context (Van Lange & Rusbult, 2014). Furthermore, Johnson and Johnson (2009) found that positive interdependence in cooperative learning can enhance individual responsibility and achievement and foster long-term group cohesion.

When it comes to pet robots, the concept of interdependence is crucial to consider. If an owner is tasked with caring for a pet robot and the robot reciprocates by providing companionship, the resultant interdependence can foster a long-term relationship. This imbues the owner with a sense of responsibility and enhances their willingness to learn with the robot. As such, this could increase the overall interaction rate between the learner and the robot, leading to a stronger relationship and potentially improved learning outcomes.

2.4 Situational Learning

Situational learning is an active learning approach that emphasizes the application of real-world experiences and practical knowledge. It was first introduced by Brown et al. (1989) and draws on the principles of apprenticeship, imitation, and practical thinking, as identified by Lave and Wenger (1991). In situational learning, learners are immersed in hands-on activities and real-life scenarios, providing them with opportunities to solve problems and apply theoretical concepts they have learned.

One of the key features of situational learning is its versatility, with researchers leveraging drama and role-playing to create scenarios that allow learners to showcase their learning outcomes. By embedding the learning material in dialogues within a structured script, learners are guided through interactive exchanges and scenarios that facilitate effective learning. For example, Erbay and Doğru (2010) used drama to have learners perform together, fostering both subject knowledge acquisition and social interaction. Similarly, Al Hakim et al. (2022a) proposed an interactive situated learning approach using a real-time feedback mechanism to guide and evaluate learners' knowledge application through interactions with robots, virtual objects, and virtual characters based on textbook context and content to enhance learning performances. By providing opportunities for learners to engage in role-play activities and real-world problem-solving, situational drama learning helps learners develop a deeper understanding of course material and cultivate particular skills.

2.5 Summary

The literature review highlights that robots have a positive impact on education, but the use of pet robots in education and research on establishing long-term relationships with them is limited. The potential benefits of utilizing pet robots in education are apparent, and this study aims to use interdependence to create a lasting relationship between the robot and the evident. Such a relationship would enhance the user's learning experience and outcomes. Finally, learners can apply their learning results by participating in situational drama role-playing scenarios with their pet robots, showcasing their ability to use the skills and gained knowledge.
3. System Design and Implementation

3.1 Learning Design

The aim of the devised learning design was not to supplant teachers' instructional approaches, but rather to employ the learning system as a means to amplify the efficacy of their teaching. Our learning design focused on fostering interdependence between learners and robots, nurturing lasting relationships and enhancing learning performance. The design integrated collaborative cooperation and mutual reliance on incentives (Johnson & Johnson, 2009; Van Lange & Rusbult, 2014), aligning learners and pet robots toward the shared objective of maintaining the pet's well-being through collaborative learning tasks that yielded food rewards. Learners, assuming roles as caregivers and trainers, engaged in interactive tasks simulating pet care duties, directly influencing the pet robot's well-being and its performance in situational dramas. In this dynamic, pet robots served as companions and performance partners, providing emotional support, companionship, and feedback. Learning materials, encompassing foundational knowledge and a dialogue-based script, expertly curated by teachers, were seamlessly integrated into a comprehensive learning script. This script served as both presentation material and guidance, to be learned by both pet robots and learners for their ultimate situational drama presentations.

3.2 Seamless Technical Design

To overcome the limitations of physical pet robots, virtual pets were introduced as a complementary solution, seamlessly integrated with the physical pet robots. They were powered by the ChatGPT API, which is an interface that allows developers to integrate the ChatGPT language model into their applications, providing advanced language capabilities, contextual understanding, and dynamic interactions. The virtual pets enabled ubiquitous communication and interaction, providing flexibility in terms of availability, capabilities, and class usage time. In the execution of the ChatGPT API program utilizing Python 3.11.3 and the GPT-3.5-turbo model, teachers effortlessly uploaded scripts in PDF, Word, or XLS formats to a designated server PC directory, establishing a custom knowledge base for the AI model.

The ChatGPT API not only produced text responses from the custom knowledge base but also played a pivotal role in instructing the virtual and physical pet robots to express emotions through facial cues and body movements (as depicted in Fig. 1). For instance, when a learner posed a question that elicited the response "I feel sad now," the API could include instructions like "Exhibit a [sad] facial expression while [nodding] in disagreement." This integration empowered the robot to not only respond textually but also embody emotions and actions, elevating interaction and immersing students in a more engaging learning experience.

Throughout the study, virtual pets accompanied learners during their out-of-class study sessions, fostering continuous engagement and interaction. Ultimately, the virtual pets seamlessly merged with the physical pet robots, allowing learners to showcase their final learning outcomes through captivating situational dramas in front of the class.
3.3 System Structure

Figure 2. System architecture of pet robot with an interdependence mechanism.

The system, built upon the foundation of the Al Hakim et al. (2023) model, comprised three integral components: a virtual pet embedded within a website (Fig. 2 [A]), a dedicated situational drama platform (Fig. 2 [B]), and a physical pet robot.

The web-based online learning platform was developed using WAMP and offered quizzes and a shopping mall. This platform empowered learners to engage in independent study outside the classroom while fostering a sense of companionship with their virtual pets. Teachers could easily insert quizzes through a user-friendly interface, customizing question content, answer choices, and feedback. Each completed quiz rewarded learners with virtual coins. The shopping mall provided an opportunity for learners to use their earned coins to purchase pet food and accessories. Through the web platform, virtual pets acted as supportive companions, offering encouragement and assistance to learners in their learning journey.

The situational drama platform to enable learners to showcase their learning outcomes through interactive drama performances. The platform is an immersive computer-based system that seamlessly connects a tablet app and the physical robot, allowing the design and performance of various learning scripts. This integration enabled learners to create interactive presentations that incorporated the physical pet robot, ultimately enhancing their learning experience and promoting their understanding of the material presented.

In this study, the ASUS ZENBO robot served as the foundation for developing the physical pet robot, establishing a tangible connection with its virtual counterpart. The developed pet model was meticulously integrated with the ZENBO SDK, harnessing its capabilities in facial expressions, body movements, audio effects, and speech module. Notably, the speech module was seamlessly integrated with the developed ChatGPT API model, enabling the physical pet robot to receive and process the user's voice, facilitating advanced language recognition and understanding. Consequently, the physical pet robot would react with a captivating array of vibrant expressions, speeches, and bodily gestures, faithfully replicating the actions and demeanor of the virtual pet. This cohesive integration between the tangible and virtual pets cultivated an exclusive and dynamic bond, enabling learners to engage in an interactive educational voyage that effortlessly traversed the confines of both the virtual and physical dimensions.
4. Method

4.1 Participants and Learning Materials

The study involved 100 university learners enrolled in Japanese for Hospitality and Tourism in Taiwan. They were divided into two groups, namely the experimental group (21 females and 29 males) and the control group (23 females and 27 males), which were further subdivided by the teacher. The groups were kept separate from each other during the experiment to ensure unbiased evaluation of the results. All participants were approximately 20 years old.

The focus of the study was on managing hospitality in Japanese restaurants that are pet-friendly. The study material encompassed several topics such as the culture of Japanese restaurants and the application of courteous expressions, all of which were incorporated into the learning script and inserted into the quizzes in the website. During the drama presentation, learners and the robot engage in a restaurant scenario. Below is an excerpt of the dialogue-based script about the culture of Japanese restaurants (translated from Japanese).

ZENBO: ("excited expression" and "spinning around") Hi there, nice to meet you!

Waitress: (surprised) Oh, a talking pet! That's amazing. Well, we're happy to have you both. Just to let you know, we do allow pets in our restaurant, but we ask that they remain on a leash and well-behaved.

Customer: Thank you for letting us know. I promise to keep him on a leash. And while we're waiting for our food, could you tell us a little more about Japanese restaurant culture?

Waitress: Of course! In Japan, it's considered polite to use honorific expressions when addressing others, especially in a formal setting like a restaurant. So instead of saying "you," we use "anata" or "omae." And it's also important to show gratitude and respect to our customers, so we always strive to provide the best service possible.

4.2 Measuring Tools

To evaluate the learners' learning outcomes, interactions, and acceptance, multiple assessment tools were employed. These included pretest and posttest questionnaires that had 15 questions each and a maximum score of 100. The teacher developed these tests and administered them before and after the experiment. In addition, database logs were used to measure interaction rate, while a five-point Likert scale questionnaire was employed in the after the experiment to measure learning acceptance. The questionnaire was based on the owner-pet relationship (McConnell et al., 2011) and learning technology acceptance (Hwang et al., 2011), modified to suit the context of the study, and consisted of 18 questions categorized into three dimensions: robot design, co-learning with the robot, and learning uses online platform. The questionnaire in this study showed high reliability with a Cronbach’s $\alpha$ value of 0.907 (N=100). The reliability of the three dimensions was confirmed by the Cronbach’s $\alpha$ values, which were 0.88, 0.90, and 0.83, respectively. Interview questions based on Al Hakim et al (2022a) were also adapted to explore factors that may impact learners’ willingness to learn with proposed approach. It consisted of six questions that were open-ended, such as "Overall, what do you think are the advantages of learning with the system?".

4.3 Procedure

The experiment spanned over six weeks and consisted of four stages, each with two hours of classes per week, followed by mandatory online practice.

- Stage I occurred during the first week and involved a pre-test, introduction to learning scripts, and an overview of the study’s system and online learning website.
During the second to fourth weeks (Stage II), learners continued to learn course materials and practiced using the online learning website independently.

Stage III took place during the fifth week and involved drama rehearsals where learners became familiar with the situational drama performance process and practiced with the robot under the teacher's guidance.

The final stage, Stage IV, was the official drama performance that occurred during the sixth week. Learners demonstrated their final learning results with the robot on situational drama platform, with each group performing independently without assistance. Post-tests, questionnaires, and interviews were conducted after the performance.

Both the experimental and control groups followed the same four-stage experimental process, were taught the same materials by the same teacher, and used online learning websites. However, the experimental group used a learning website with a virtual pet robot that required learners to obtain virtual coins during online practice to purchase food and take care of the virtual robot's cages. The control group used a website without a virtual pet robot for learning. During the drama performance, the experimental group used a physical robot with pet characteristics that displayed lively expressions and actions according to the script, while the control group used a general robot that only read script lines and did not have pet functions.

5. Findings

5.1 RQ1: Learners’ Learning Outcomes

The analysis employed analysis of covariance (ANCOVA) to compare the post-test scores of the experimental and control groups, with the pre-test scores used as a covariate. The homogeneity test of intra-group regression coefficient revealed no significant interaction between the two groups and the pre-test scores \((F = 0.42, p = 0.29 > 0.05)\), allowing the use of pre-test scores as a covariate. Subsequently, ANCOVA was utilized to analyze the post-test scores of the two groups. The results of the statistical analysis, as displayed in Table 1, indicate a significant difference in the mean scores between the experimental and control groups \((F = 2.19, p = 0.01 < 0.05)\). This suggests a notable disparity in the post-test scores between the two groups.

Table 1. ANCOVA result of learners’ learning outcomes

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Adjusted mean</th>
<th>F</th>
<th>(\eta^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental</td>
<td>50</td>
<td>80.98</td>
<td>4.89</td>
<td>81.03</td>
<td>2.19*</td>
<td>0.07</td>
</tr>
<tr>
<td>Control</td>
<td>50</td>
<td>76.21</td>
<td>6.89</td>
<td>75.46</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The findings support the claim that a pet robot equipped with an interdependence mechanism is a more effective tool for improving learning outcomes compared to a conventional robot. These results align with earlier investigations (Chen, 2012; Lin et al., 2022) that have demonstrated the positive impact of including a pet-like interaction mode in the learning system, which enhances learners learning achievement. By taking care of their virtual pets, learners felt a sense of responsibility, knowing that neglecting their pet would result in its demise, thereby encouraging them to learn and earn virtual coins to buy supplies. Assigning each learner an interdependent pet can, therefore, foster a sense of duty towards the pet and their own learning, ultimately leading to improved learning outcomes.

5.2 RQ2: Interaction Rates

The average daily interaction in learning activities was recorded for both the control and experimental groups across four stages. Results show that the experimental group had significantly higher activity levels than the control group (see Fig. 3). This suggests that the
presence of the pet robot served as a stimulus for the learners to engage more actively in the learning process. This phenomenon is attributed to the reciprocity of pet care (Lawton, 2017; Chen, 2012), where pupils engage in more interactions to take care of their pets by feeding and playing with them. In this study, learners increase learning frequency to earn coins and purchase pet food due to the need to take care of the pet robots, while pet robots provide feedback and support to learners at all times, forming a positive interactive relationship that keeps learners engaged in learning activities. Notably, based on the Figure 2, it was observed that the interaction rate among the experimental and control groups continued to increase in the final stage, that is official drama performing. This might attribute to the motivation-driven effect (Al Hakim et al., 2022b), which compelled learners to study before the final exhibition with the robots to be displayed to classmates and teachers.

**Figure 3.** The average daily interaction in learning activities among groups.

5.3 RQ3: Learners’ Acceptance of Learning

The responses to the questionnaire were analyzed with independent *t*-test to compare the learners’ acceptance of learning after the treatment. As shown by Table 2, in terms of the design perception of the robot, both groups interacted with the physical robot and had similar perceptions, with no significant difference found (*p* = 0.620 > 0.050). It means that both of groups felt that the appearance, expression, voice, and interaction of the developed robot models, whether it has pet feature or not, were similar. In terms of co-learning with the robot, both groups scored higher than the average level, with the experimental group having a significantly higher average score (*M* = 4.130) than the control group (*M* = 3.850), indicating that pet robots can provide a better educational encounter for learners. On the other hand, there was no significant difference between the two groups in terms of user acceptance when using the online learning platform with or without the pet robot function (*p* = 0.220 > 0.050).

**Table 2.** *T*-test result of learners’ acceptance of learning

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robot design acceptance (e.g., <em>I find ZENBO’s appearance and expressions adorable, like a pet</em>)</td>
<td>Experimental</td>
<td>3.70</td>
<td>0.76</td>
<td>0.253</td>
<td>0.617</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>3.61</td>
<td>0.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction of co-learning with the robot (e.g., <em>Having ZENBO as a companion has helped me learn the materials better</em>)</td>
<td>Experimental</td>
<td>4.13</td>
<td>0.66</td>
<td>8.91</td>
<td>0.004*</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>3.85</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction of using the online platform (e.g., <em>The online learning platform has enhanced my learning experience</em>)</td>
<td>Experimental</td>
<td>3.61</td>
<td>0.45</td>
<td>1.523</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>3.60</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p* < 0.05
Furthermore, to gain a better understanding of how learners perceive the proposed learning approach, ten-minute interviews were conducted with six learners from the experimental group, representing high, medium, and low academic achievement. The outcomes of these interviews revealed intriguing findings. Firstly, learners expressed their fondness for the pet robot's cute appearance and lifelike behavior, finding them highly appealing. Motivated by a sense of responsibility towards their virtual pet and a genuine fear of it going hungry, learners engaged with the robot more earnestly. They also highlighted the seamless connection between the virtual and physical pet robots, enabling them to extend their emotional bond from the virtual realm to the physical world. However, some interviewees suggested that incorporating more pet-like facial expressions and personalized interactive modes could greatly enhance the enjoyment factor of interacting with pet robots. Specifically, they expressed their desire for interactive behaviors like cuddling and spinning, which would further enrich the design and interaction possibilities of the pet robot.

These insightful interview responses indicate that the proposed learning approach not only captivates learners by leveraging the novelty of the robot but also fosters motivation through a heightened sense of responsibility towards their virtual pets. Furthermore, learners expressed their anticipation for more design options and personalized responses from their pet robots, reflecting their eagerness to have a unique and customized pet robot that builds upon the established relationship.

6. Conclusion, Drawbacks, and Further Research

This study presents a novel approach to educational robot design by integrating interdependence theory. The aim is to develop pet robots that establish enduring relationships with learners, thereby enhancing their learning performance. The interdependent pet mechanism, encompassing the owner's responsibility for the pet's care and the pet's provision of emotional support, is seamlessly integrated into both virtual and physical platforms, along with situational drama learning presentations. This integration empowers learners to take care of their pet robots anytime, anywhere, fostering a sense of responsibility for their own learning.

The results of the experiment indicate that robots designed as interdependent pets have a more positive impact on learners in terms of learning outcomes, long-term interactions, and acceptance of learning than conventional robots. The study provides a new insight on the use of robots in education, combining generative AI, mobile and reality technologies, and practices, resulting in an intelligent learning environment where robots act as interdependent pets. Educators and system developers can incorporate this approach into the learning system to create a highly engaging and interactive learning experience for learners, enabling them to establish lasting relationships with learning agents and improve their learning performances.

However, this study was limited conducted on a Japanese for Hospitality and Tourism course for university learners, making it uncertain whether the findings can be generalized. Further experiment could explore the potential of this approach for other age groups and subjects. Also, future studies should consider several potential enhancements to pet robot design based on valuable feedback received from learners. Firstly, there is a need to expand and enhance the range of pet-like features beyond basic interactions such as satiety and basic expressions. Incorporating a broader spectrum of pet emotions, improved cleanliness, and more specific behaviors and attire could potentially increase user interest and willingness to engage with the pet-like robots. Secondly, strengthening the connection between virtual and physical robots can be achieved by assigning unique names, distinct speech patterns, and individual personalities to each pet, ultimately would enhance the personalization and deepen the connection between the virtual and physical manifestations of the robots.

Acknowledgements

The study was supported by the National Science and Technology Council, Taiwan. Grand Number: MOST 110-2511-H-008-004-MY3 and NSTC 112-2811-H-008-006.
References


Calvo-Barajas, N., & Castellano, G. (2022, September). "I have an idea!" enhancing children’s verbal creativity through repeated interactions with a virtual robot. In *Proc. of the 22nd ACM Int. Conf. on Intell. Virtual Agents* (pp. 1-8).


Analysis to Creation: Using the ADDIE Model to Develop an Educational Game for Children

Yufan ZHANG\textsuperscript{a}, Nurul Nadwa ZULKIFLI\textsuperscript{b} & Ahmad Fauzi MOHD AYUB\textsuperscript{c},
Zewen SHANG\textsuperscript{d}
\textsuperscript{a}Yufan ZHANG, Universiti Putra Malaysia, Malaysia
\textsuperscript{b}Nurul Nadwa ZULKIFLI, Universiti Putra Malaysia, Malaysia
\textsuperscript{c}Ahmad Fauzi MOHD AYUB, Universiti Putra Malaysia, Malaysia
\textsuperscript{d}Zewen SHANG, Singapore University of Technology and Design, Singapore
\textsuperscript{*} nurulnadwa@upm.edu.my

Abstract: Creating educational games for children is a notable emphasis, especially when using the ADDIE model as a game development framework. Furthermore, we introduce the ADDIE model, a recognized instructional design framework, to illustrate its potential in the video game industry. In this study, by using constructivist and social learning theories, educational play can be fostered through exploration, cooperation, and reflection. Hence, we create a children's educational video game using the ADDIE model to explore creating suitable educational games. Striking the right balance between education and entertainment necessitates a comprehensive grasp of how children learn and develop. By leveraging this knowledge, developers and educators can create video games that are both enjoyable and promote learning and development in children.

Keywords: ADDIE model, children's educational game, game design

1. Introduction

In response to the growing recognition of digital games for early childhood learning, we developed an educational video game for children under 7, focusing on reading, numeracy, social, and emotional skills. Our study is utilizing the ADDIE paradigm, a systematic instructional design framework, we detail our game’s creation in this framework.

2. Related Video Game Analysis


3. Applying the ADDIE Model for Effective Game Development

3.1 ADDIE Model

Game designers employ ADDIE to create well-structured, engaging games aligned with objectives for the intended audience (Davis, 2013; Muruganantham, 2015). As shown in Figure 1, the ADDIE paradigm offers a structured, collaborative approach to game
development, facilitating clear goal setting, comprehensive planning, and accurate progress evaluation.

Figure 1. The ADDIE Paradigm

3.2 ADDIE Model in Game Design Process

The ADDIE model ensures purposeful game development that meets audience needs, as seen in Braad et al. (2016) study. It offers designers to create effective and engaging games as illustrated in Figure 2.

Figure 2. The ADDIE Game Design Process

4. The Production and Design of Educational Games for Children

In this research, we present an educational game flowchart in Figure 3. The flowchart details the different actions and learning objectives of the game.

Figure 3. The Game Process
4.1 Tool Selection for Development

We present our decision to develop our app on the iOS platform instead of Android. There were several factors that influenced this choice, including the relative ease of development and testing on the iOS platform, and the strong user engagement.

4.2 Function Realization

4.2.1 The Realization of Game Story

As our primary target audience is comprised of young children, specifically those aged below seven, our game necessitates a theme that is both user-friendly and possesses a more comprehensive storyline, as opposed to a complex user interface.

4.2.2 The Realization of User Interaction

As our main audience is under seven years old, we’ve connected custom buttons to visually appealing elements like tables and classrooms, creating a vibrant and captivating user interface. Utilizing the ADDIE model ensures clear learning goals, content alignment, and mechanics congruence (Kapp, 2012; Lim et al., 2013). Designing personalized educational games for children, aligned with learning objectives, and employing frameworks like ADDIE, constructivism, and scaffolding, can result in immersive and impactful learning experiences.

Acknowledgements

I extend heartfelt gratitude to my tutors, Dr. Nadwa and Professor Fauzi, whose unwavering guidance and support have been instrumental in my studies. Special thanks to a fellow PhD student for their invaluable technical support alongside Zewen during this research.

References


Case study-based research on understanding app user engagement to develop environmental literacy of urban residents

Ewa DUDA
Maria Grzegorzewska University, Poland
eduda@aps.edu.pl

Abstract: Environmental literacy enables action to be taken against climate change. These actions should naturally be supported by developing information technology. Mobile applications can therefore be successfully used as a tool for learning, awareness-raising, but also for shaping environmentally friendly behavior. This article presents an analysis of the functionalities of mobile apps that enhance or undermine user engagement with the application. Qualitative research, carried out as a case study, was conducted by analyzing publicly available user reviews. The results show that a well-designed application can provide valuable support for ongoing educational interventions aimed at enhancing the environmental literacy of urban residents. It presents indications that are worth using when designing environmental applications to ensure a high level of future user engagement. Results can be successfully used by mobile app designers, but also educators, city authorities, or local activist.

Keywords: environmental literacy, educational applications, green transformation, pro-environmental behavior, urban education, user engagement

1. Introduction

Developing the environmental literacy of urban residents is becoming a necessity in the face of ongoing climate change. Such literacy is perceived as (1) understanding the resources at our disposal, (2) understanding that the actions of people (social beings) have roots and consequences, (3) the ability to work together to achieve common, rational goals, (4) the ability to responsibly obtain natural goods for basic needs and desires (Roth, 1992). Indeed, a pro-environmental approach must be embraced at multiple levels, taking care not only to strengthen awareness but also knowledge, attitudes and practical skills that will be an integral part of urban residents’ conscious daily routines.

The changing society needs and the rapid development of information technology imply the question of how to effectively encourage people to engage in a process of shaping their environmental literacy, and to maintain that engagement over time. This study aims to explore the issue of co-development of solutions to foster green social transformation. The initial concept of the tool to support educational interventions among urban residents (Duda et al., 2022) is based on the use of a mobile application to facilitate environmentally friendly behaviors. To address this, the study attempts to answer the following research question:

- What features of a mobile application for environmental literacy encourage users to adopt environmentally friendly behavior?
- What features of a mobile application for environmental literacy discourage users to adopt environmentally friendly behavior?

2. Theoretical Framework

2.1 Environmental Literacy
The term environmental literacy is understood as an advanced level of conscious human interaction with the environment. It is defined as “the ability to appropriately read and to utilize environmental information, to anticipate rebound effects, and to adapt according to information about environmental resources and systems and their dynamics” (Scholz, 2011:8). On the other hand, Cole drew attention to the cultural aspect of the concept, where “environmental literacy can be understood as a culturally specific body of knowledge that fosters particular ways of thinking and acting in the world” (Cole, 2007:39).

The development of environmental literacy is based on active education that goes beyond the mere transmission of information. The vast majority of literature on developing environmental literacy focuses on formal, informal or non-formal education (Forsyth, 2018; Taylor, 2020). Research on aspects related to the use of mobile apps in education for environmental literacy is much more limited. Most often, they focus on application-specific research such as educational process among participants and gamers (Bowser et al., 2014), motivators to reinforce pro-environmental behavior (Pacheco, & Faria, 2022).

2.2 User Engagement

The definition used in this study is as proposed by Attfield, Kazai and Lalmas (2011). Regarding this definition “user engagement is the emotional, cognitive and behavioural connection that exists, at any point in time and possibly over time, between a user and a resource, […] user engagement with a technological resource is not just about how a single interaction unfolds, but about how and why people develop a relationship with technology and integrate it into their lives” (Attfield, Kazai, & Lalmas, 2011:2).

3. Methods

3.1 Procedure

The Google Play Store platform was used to search for mobile applications. Downloadable, English-language mobile applications with publicly available user reviews were included in the searches. The keyword used was "environment", next the search continued using the snowball method. The apps were examined for their widest range of features addressing environmental issues. The first search yielded 86 results. In the next step entertainment games, dictionaries, and quiz applications were excluded. Finally, three apps most related to the research objective were selected, with a significant number of user reviews available. They were: “Environment Challenge” (106), “Earth Hero: Climate Change” (445) and “JouleBug” (286 available reviews).

3.2 Data Analysis

As the aim of the study was to explore constructs and categories emerging from both positive and negative user statements, the study employed inductive analysis (Mary, & Pour, 2022). In this study, the data was divided into two main thematic areas. The first comprised functionalities that, according to users, stimulate them to undertake environmental behavior through the app. The second covered functionalities that discourage the use of the app, thus not encouraging environmental efforts.

4. Results and Discussion

Analyzing the posted reviews, a reflection arises that the users of the above applications are people who are interested in environmental problems, but their knowledge in this area is not very advanced and they need regular stimulation to take environmentally friendly actions. The app makes it easier for people to be more eco-friendly, especially those who may be struggling with environmental anxiety during the climate crisis.

Users stressed that their lifestyles have changed a bit since they started using the app. The friendly, non-judgmental tone of the notifications helped them. Also positive was the lack
of competition, a system based on goal-setting and reminders. Tips and suggestions were easy to follow and incorporate into a daily routine. A higher rating was given to the app providing measurable challenges and more personalized, as is in line with Morreale et al. (2015) research. Challenges that were too general or those that did not provide a point of reference caused a decline in engagement. Users questioned issues such as having a challenge to consume some percent less power when they have no idea about the quantity of their current consumption. The challenges formulated should therefore be more flexible.

Also demotivating for some users were vague statements about the challenges. While users received daily reminders to keep up with the challenges, they lacked detailed guidance on how to do this. On the other hand, seven out of ten users were satisfied with the support provided to them. They enjoyed the articles about the issues in the environment and steps to minimizing human impact. The data provided and the concept of environmental challenges was both useful and educative. They declared, that app made them more aware of their day to day actions and should be further developed and maintained. Such applications can help to develop environmentally friendly habits, and increase users' knowledge (Rosal et al., 2021).

5. Conclusion

The main novelty of the study is based on the uniqueness of the topic addressed. In addition to the small number of studies on the formation of various pro-environmental habits and behaviors of residents, it is difficult to find research on the engagement of mobile app users, especially based on a qualitative analysis of the opinions of current app users. This study, therefore, extends the available literature on the subject by contributing to the promotion of environmental action. The results of the research support the application design process of the project in which the authors of the text are involved. They can be also successfully used by other mobile app designers, educators, and city authorities. A limitation of this study is that it was carried out by a single researcher, which may introduce a bias in the coding process.

Acknowledgements

The research is supported funding from Iceland, Liechtenstein and Norway under the EEA Funds and the state budget of Poland via the National Centre for Research and Development within "Applied Research" Programme, grant no. NOR/IdeaLab/GC/0003/2020-00.

References

Design and development of a game to improve self-efficacy: A case study of addressing modes learning

Fuzheng ZHAOa, Danqing LUOa, Etsuko KUMAMOTOb & Chengjiu YINC

a Graduate School of System Informatics, Kobe University, Japan
b Information Science and Technology Center, Kobe University, Japan
c Research Institute for Information Technology, Kyushu University, Japan

*yin.chengjiu.247@m.kyushu-u.ac.jp

Abstract: The low self-efficacy has been an important issue in the design of instructional methods, and its negative effects are mainly shown by the reduced willingness to learn and the high dropout and failure rates. Game-based learning has received attention in the design of t instructional methods because of the combination of learning content and games. In this study, we use addressing modes learning in operation system course as a case object to explore the potential of improving the low self-efficacy by game-based learning. To this end, a side-scrolling video game was designed to complete the case study task to examine the effect of game-based learning on solving the low self-efficacy issue. According to the experimental results, there was no difference in learning achievement between the experimental group students who used game-based learning, and the control group students who used traditional learning methods. However, with similar scores in learning achievement, the experimental group students showed higher self-efficacy than the control group students. In addition, it was shown that the game-based learning approach did not impose additional cognitive load.

Keywords: Game-based learning, self-efficacy, addressing modes learning

1. Introduction

Self-efficacy in learning activities has long been of interest to researchers. Since this issue is closely related to students’ learning achievement and psychological cognitions. For example, appropriate self-efficacy can help students reduce learning stress, increase learning efficiency and maintain interest and engagement in learning activities (Silverman & Casazza, 2000). Conversely, the negative effects of low self-efficacy are obvious. Students who spend long periods of time in low self-efficacy face higher levels of studying stress, exhibit avoidance behavior and tendency to boredom, which ultimately leads to high dropout rates (Martin & Marsh, 2003).

High dropout and failure rates in computer-related courses are a common problem (Gomes & Mendes, 2007), and Heward (2003) argues that students who face challenging and unfamiliar learning content are more likely to develop low self-efficacy, reducing motivation and willingness to learn. Meanwhile, Weina, Ping, & Shuai (2018) confirmed that the main challenges in computer courses come from abstracted concepts where knowledge points are dispersed and the intrinsic logic relationship between concepts is hard to find. Therefore, instructional methods that match the learning content receive attentions.

The conceptualization, design principles, and applications of game-based learning have gained great attention (Tobias, Fletcher, & Wind, 2014). Game-based learning is not a new thing, especially for contemporary students who are accustomed to electronic devices early on. In this study, addressing modes learning in operation system course is used as a case object to use game-based learning as an attempt to solve the issues of low self-efficacy that arises in computer-related courses.
2. Literature review

2.1 Addressing modes learning in operation system course

Operation system course is a core foundation course with many concepts, strong principles, and high degree of abstraction (Weina, Ping, & Shuai, 2018), which related to much hardware knowledge. As a result, high drop out and failure rates have been reported with focusing on much attentions (Gomes & Mendes, 2007). Addressing modes means the specification of data locations required by an operation. Also, mastering addressing modes involves thinking and reasoning that usually required for operation system learning. As one of difficulties in learning operation system, more demands for practicable lessons with reasonable learning approaches than purely theoretical ones have been made when learning addressing modes (Bolanakis et al., 2008). Faced with challenging learning content, it is more likely to have insufficient self-efficacy (Martin & Marsh, 2003).

2.2 Impact of low self-efficacy on learning activities

Self-efficacy theory originally stemmed from objective analysis on changes achieved in fearful and avoidant behavior. Bandura described self-efficacy is concerned with individuals’ beliefs in their control over challenging demands (Bandura, 1994). Diverse effects are produced depending on various levels on self-efficacy. It was found that a strong sense of self-efficacy is highly correlated with improvement of human accomplishment (Maddux & Gosselin, 2012). Schunk (1995) reported that people with high self-efficacy regard difficult tasks as challenges to be mastered rather than as threats to be avoided. Intrinsic interest and deep engrossment in learning activities may be fostered easily with help of such self-efficacy (Zajacova, Lynch & Espenshade, 2005). Also, higher self-efficacy may reduce learning frustration, facilitate learner’s effort and improve sustaining interest in challenging activities (Silverman & Casazza, 2000). Although self-efficacy has moderated relationship with learning performance, it is not necessarily that high self-efficacy leads to higher learning performance (Balkis, 2011). Davis (2009) found that people with higher self-efficacy have higher subsequent learning performance.

2.3 Challenges on game-based learning design

Application potential of game-based learning was ground in a mount of research. The acceptance of game-based learning has been widely taken in educational context. Among application benefits, facilitating learning by fostering multiple perspectives such as cognitive, affective, behavioral and sociocultural engagement is full of expectation from actual application (Jan et al., 2015). However, emphasis only on learning performance has been fading, in contrast to growing shift to a combination of learning and psychology effect. Notably, some questions were raised along with attention on how to make use of the game promise for education purposes, while taking balance of entertainment and learning in account.

Another challenge of converting learning objectives to game design is prominent. When designing game based on particular pedagogy that requires logical ability to understand abstractness, it will exacerbate the challenge of this conversation (Moreno-Ger et al, 2008). Pedagogical requirements driven by difficult learning content may result in low willingness to persist involvement in learning tasks (Westwood, 2003). General learning difficulties can happen owing to inadequate or inappropriate teaching (Westwood, 2003), especially in the early stages of learning with complicated learning patterns that hard to determine reasons (Zhao, Hwang, & Yin, 2021).

Game designing lacks of consensus on uniform rules, because it involves of complicate game-design factors such as learning objectives, mechanism, fantasy value, interaction, freedom, narrative, sensation, challenges, sociality, and mystery (Shi & Shih,
2015). One of the most critical principles was presented by Shabalina et al. (2014), emphasizing that a sequence of game activities is structured by coming learning objectives and game activities. To support the effect embedding game in learning, meeting the needs of course with a focus on learning goals, feedback, and interaction in the game functions (Alaswad & Nadolny, 2015). So that, it is essential to consider learning attributes included of above game-design factors and game functions.

The studying questions are as follows:
- How to combine learning contents and game operational processes in game design?
- Could the developed game improve self-efficacy and academic achievement?
- Could the developed game bring additional cognitive load in solving low-efficacy issues?

3. Game system design and development

The game module includes (1) UI module, consists of navigation mode, content display, and basic controls. (2) Physics module, the game needs to interact with the performance of objects given physical characteristics, such as the size of the force when the character jumps. (3) Scene module, the terrain, level and character management in the game. When developing the game, the three modules were developed using the Unity game engine as the development tool. Firstly, use the physics system, graphics system and audio elements in Unity to develop scene module. Subsequently, in the built game scene, game characters are added and the character behavior and game logic in the scene are implemented through Unity scripts. Finally, use uGUI to design the game interaction interface, such as game effects. Next, the following two aspects will mainly be introduced: how to translate learning content into game elements and design core game modules.

3.1 Implement of learning content in the design of game system

Concepts and operational processes are the two main learning content when learning addressing modes. These can be further refined into: data locations, proper nouns and addressing operations. In game design and development, three learning content are translated into game maps, game terms and game activities respectively.

3.1.1 Correspondence between data locations and game maps

The storage and processing of data during addressing involves three main data locations: the CPU (processor), the memory and the auxiliary memory (hard disk). Firstly, there are direct and indirect ways of accessing data. CPU has direct access to memory data and no direct access to hard disk data. This indirect access to hard disk data by the CPU can only be done with the help of memory. Secondly, the method of data storage shown in Figure 1. Memory and hard disk data storage relies on a form of data storage called data blocks. Unlike hard disks, memory data storage also requires a form of data storage known as paging and segmentation tables.

![Figure 1. Virtual memory management.](image)

Based on computer structures and data locations, we designed game maps relatively to gamify the above knowledge. Figure 2 represents the corresponding relationship between
the data locations and game maps. In Figure 2, the areas where game character movement and game activities occur in the game correspond to the locations where the data storages. For example, the memory map in the game is the memory location where the data storages. Also, the detailed knowledge content required for the different data locations during addressing is matched to the corresponding map. Specifically, bytes are represented by cells in a memory map, and page and segment tables are represented by cell and row tables respectively. In addition, gates are set up between different game maps, and when the character enters the gate of a different map this indicates the corresponding data storage location for reading.

![Figure 2. Correspondence between data locations and game maps.](image)

### 3.1.2 Correspondence between proper nouns and the design of game items

Main challenges in game design come from transforming the proper nouns in OS into the corresponding game items. In Figure 3, the game clues and props are matched to the OS proper nouns. First, game clues. The initial clue of the game is the logical address stored in CPU, since the logical address does not need to be calculated as the initial condition for addressing. The second game clue is obtained by assigning a row in the page or segment table from the logical address, which is represented by a prop box with a clue marker. Secondly, game props. According to the addressing process after the physical address is calculated, the correct block in memory needs to be read. Doors in the game represent the block, and the position of the door indicates the number of blocks. Once inside the door, the character can search for specific data in that block map, represented by a prop box.

![Figure 3. Correspondence between proper nouns and game items.](image)

### 3.1.3 Correspondence between addressing operations and game actions

The transformation of proper nouns in OS into game items is based on addressing process, where core information about the operational process in addressing is gamified into game clues and props.
The figure 4 shows that the start of addressing operation in the game is represented by the character accepting a task from the task board (a) and obtaining new clues through (b) and (c). Entering the map (e) where data block is stored in prop box through the door (d), finding the corresponding pro box (f) according to the calculation results and completing the storage of the items (g) and (h).

3.2 Learning strategy incorporated in the design of system modules

As shown in Figure 5, three game modules were designed. 1) mission module contained cues to help students understand the sequence of steps in the addressing modes, 2) guide module scaffolds participants with instructions on how to operate the game and learn the task (Yin et al., 2013), and 3) feedback module was responsible for feedback on task status and character position.

![Figure 5. System models.](image)

3.2.1 Game clues in mission module

In the mission model, we gamify six addressing modes and design them into six missions. As shown in Table 1, the six missions correspond to the six addressing modes. In each mission, the game provides the corresponding operational steps for the specific addressing mode. The Figure 6 shows that a specific operation process about mission 1.

First, the participant gets the first clue from the task board, which is the logical address stored in the CPU. Next, participants were given a second clue in the page table prop box, which was to calculate the physical address by reading the page table. In the task, \( v=1 \) means that a block of data in memory is read, otherwise it means that a block of data on the hard disk needs to be read. When \( v=1 \), the participant calculates the physical address and reads a block of data in memory in the next step. The participant then reads the data in a block based on the calculated offset, while the character opens a box in the block graph. Finally, the CPU gets the data and the game task gets completed.
Table 1. *Six missions about addressing modes*

<table>
<thead>
<tr>
<th>Mission number</th>
<th>Addressing modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mission 1</td>
<td>Obtaining data in memory using page table.</td>
</tr>
<tr>
<td>Mission 2</td>
<td>Obtaining data in hard disk using page table without swapping in.</td>
</tr>
<tr>
<td>Mission 3</td>
<td>Obtaining data in hard disk using page table with swapping in.</td>
</tr>
<tr>
<td>Mission 4</td>
<td>Swapping out data in hard disk.</td>
</tr>
<tr>
<td>Mission 5</td>
<td>Obtaining data in hard disk by segment table without error reporting</td>
</tr>
<tr>
<td>Mission 6</td>
<td>Obtaining data in hard disk by segment table with error reporting</td>
</tr>
</tbody>
</table>

*Figure 6. Reading data process in memory by paging.*

### 3.2.2 Tips in feedback module

This module includes game operation and learning mission instruction, as shown in Figure 7. The game operation guidance mainly instructs game participants how to manipulate character behavior, switch between different maps and interact with elements in the map. The learning mission guide can take on the role of an instructor, guiding game participants on how to acquire tasks, showing the meaning and function of game elements, and suggesting solutions for mission completion. Notably, learning missions are designed to provide scaffolding for game participants with timely guidance, which decreases as they progress through the missions and improve their learning abilities. When game participants perform the mission for the first time, they are provided with detailed instructional information. As the number of games increases, the instructional information decreases.

*Figure 7. The mission guidance.*
3.2.3 Tips in guide module

This module is designed to keep game participants with the mission status, providing information on game location and clues in a timely manner. As shown in Figure 8, the mission location information can help game participants understand the progress changes of the mission and facilitate mastering the current mission. The participant location information, which is a kind of clue information being fed back, can increase the understanding of the game participants on the updated status of the game clues and props.

![Figure 8. The feedback information.](image)

4. Experiment design

4.1 Participants

An experiment was conducted to verify the contributiveness of the developed game in terms of learning effectiveness and self-efficacy.

A total of 39 participants, from a computer-related discipline at a Japan university, took part in the experiment. The experiment participants were assigned to an experimental and a control group. The participants in the experimental group learned the addressing model using a game-based learning approach, while the participants in the control group completed the learning task by reading the textbook.

4.2 Instruments

The instruments involved in the experiment mainly included measures of learning performance, self-efficacy, and cognitive load. The learning performance measures consisted of pre-test and post-test. The pre-test consisted of 5 multiple-choice and 9 short-answer questions, and its measure was the basics of operating systems. The post-test, on the other hand, consisted of nine comprehensive questions limited to knowledge of the search model. Cognitive load test is a way to examine the impact of a learning strategy on students’ learning activities (Zhao et al., 2023). Self-efficacy and cognitive load measures were administered through a questionnaire modified from the questionnaire developed by Pintrich and De Groot (1990).

4.3 Procedure

Figure 9 shows the experiment procedure. Before the learning activity, the experimental and control groups took a 3-minute pre-test and a 10-minute pre-questionnaire, respectively. Subsequently, both of two groups were given a 30-minute learning activity at the same time, in which the experimental group used the designed game to learn the addressing mode, while the control group learned by reading the textbook. Afterward, a 10-minute post-test and a 10-minute post-questionnaire were asked.
5. Experiment results

After data processing, we obtained valid data for 34 participants. To verify the contribution of the designed game in learning achievement, self-efficacy, and cognitive load. First, a normal distribution test was adopted to analyze three kinds of data using the Shapiro-Wilk test. Specifically, the post-test showed significance (w=0.928, p=0.027), implying that it does not meet the assumption of normality (p<0.05). In addition, pre-test (w=0.980, p=0.775), self-efficacy (w=0.972, p=0.529), and cognitive load (w=0.952, p=0.142) did not present significance (p>0.05), implying that it meets the assumption of normality. So that, a t-test was used to analyze whether there were differences in pre-test, self-efficacy, and cognitive load. Also, the non-parametric Mann-Whitney Test was used to analyze whether there was a difference in post-test.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>Experiment</td>
<td>22</td>
<td>33.770</td>
<td>6.414</td>
<td>0.607*</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>31.420</td>
<td>5.838</td>
<td></td>
</tr>
</tbody>
</table>

Note. *p > .05.

The game learning method on learning performance was examined using the Mann-Whitney test. The table 3 shows that the mean rank of the experimental and control groups were 18.430 and 15.790, respectively, and the sum of ranks were 405.500 and 189.500. Based on Mann-Whitney results (z = -0.751, p>.05) indicate that students in the experimental group who used the game-learning approach to learn addressing model did not differ significantly from the control group students in terms of learning performance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>Experiment</td>
<td>22</td>
<td>18.430</td>
<td>405.500</td>
<td>-.751*</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>15.790</td>
<td>189.500</td>
<td></td>
</tr>
</tbody>
</table>

Note. *p > .05.

The t-test was used to examine the differences in self-efficacy between the experimental and control groups. The table 4 shows that the Mean was 20.41 and 15.33 for the experimental and control groups and the S.D. was 3.763 and 5.416, respectively. The t-test results (t = 3.213, p < .05) show that students in the experimental and control groups were statistically different in terms of self-efficacy. In other words, students who used the game-learning addressing model had higher self-efficacy than had higher self-efficacy.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Sum of Ranks</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experiment</td>
<td>22</td>
<td>20.41</td>
<td>405.500</td>
<td>3.213*</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>12</td>
<td>15.33</td>
<td>189.500</td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Group</td>
<td>N</td>
<td>Mean</td>
<td>S.D.</td>
<td>t</td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>----</td>
<td>-------</td>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>Pre-test</td>
<td>Experiment group</td>
<td>22</td>
<td>20.41</td>
<td>3.763</td>
<td>3.213*</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>12</td>
<td>15.33</td>
<td>5.416</td>
<td></td>
</tr>
</tbody>
</table>

Note. *p <.05.

According to table 5 shows that the experimental and control groups in Mean were 23.50 and 21.92, respectively, and the S.D. was 5.570 and 4.680. The statistical results (t = 3.213, p>.05) show that the students in the experimental group with the game-learning addressing mode had similar cognitive load as the students with the traditional approach, which indicates that the game-learning approach did not bring more cognitive stress to control group students.

Table 5. Descriptive statistics and t-test results of the cognitive load

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>Experiment group</td>
<td>22</td>
<td>23.50</td>
<td>5.570</td>
<td>-.835*</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>12</td>
<td>21.92</td>
<td>4.680</td>
<td></td>
</tr>
</tbody>
</table>

Note. *p > .05.

6. Conclusions

In the study, a game was developed to solve the low self-efficacy issue of the students in the learning activity by combining the learning content and the game steps with the design concept of game learning. Besides, the game was designed by taking the addressing mode of the operating system course as the learning content in three ways: using the three data locations of hard disk, memory and CPU as the game map; transforming the per nouns such as block, data and offset into one-to-one game elements; converting the operation behaviors in the addressing mode into game operation behaviors. Also, the game is designed with three modules that contain learning strategies. They are the mission module, which provides cues; the feedback module, which provides tips; and the guidance module, which has a similar functions with giving tips.

To examine whether the developed game can solve the problem of low self-efficacy exhibited by the students in the learning activities, a comparison experiment was conducted. Based on the results of the experiment in three dimensions: learning achievement, self-efficacy, and cognitive load, the game was statistically analyzed for avoiding low self-efficacy behavior. First, according to the results of the t-test, the students in the experimental group who used game learning, did not differ significantly from the students in the control group who used traditional learning in terms of learning achievement. Second, with similar learning achievement, students in the experimental group differed significantly from students in the control group in terms of self-efficacy, and students in the experimental group had higher self-efficacy, which implied that the developed game could improve self-efficacy. In addition, T test was adopted to test whether the introduction of the game learning method brought about cognitive load on students. Based on the results showed that there was no significant difference between the experimental and control groups in terms of cognitive load. This indicates that the game-based learning strategy did not bring about additional cognitive load.

Acknowledgements

This research was partially supported by Grants-in-Aid for Scientific Research No. 21H00905 from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan.
References


Investigating Cognitive Biases in Self-Explanation Behaviors during Game-based Learning about Mathematics

J.M. Alexandra ANDRESa*, Elizabeth B. CLOUDEb, Ryan S. BAKERb & Seiyon LEEb
aGraduate School of Education, University of Pennsylvania, USA
*bandres@upenn.edu

Abstract: The study investigates the impact of cognitive biases on middle-school students’ affective experiences while learning about math in a game-based learning environment (GBLE). The study focused on students’ confrustion, an affect construct that unifies the several manifestations of confusion and frustration. We studied confrustion in the context of students self-explaining erroneous examples, where they had to find and fix common errors in given math problems and self-explain their problem-solving processes either with or without scaffolding. Text replays were utilized to examine student interactions during game-based learning and identify behaviors that emerged in response to cognitive biases and affect and its impact on learning and performance outcomes. The results revealed that students who demonstrated more pseudo-confidence in their self-explanations had higher self-reported self-efficacy, but were more likely to submit incongruent responses, exhibit confrustion, make errors, and take longer to finish the game. Overall, the findings show that students were vulnerable to cognitive biases and did not always respond in ways that accurately reflected their approach to solving math problems. The insights into how students approach and learn from math games inform the design and implementation of GBLEs by addressing cognitive biases.

Keywords: game-based learning, confrustion, cognitive bias, confidence, self-explanation, self-efficacy

1. Introduction

Game elements designed to scaffold instructional support are essential for promoting behaviors that support mathematical learning (e.g., prompting students to generate self-explanations about their solutions; Ke et al., 2019; Wouters & van Oostendorp, 2013). A study by McLaren and others (2022) showed that students who generated open-ended and focused self-explanations (i.e., open-ended responses answering a focused prompt), compared to menu-based self-explanations (i.e., select an explanation from a menu), had significantly higher learning. However, in further studies, Johnson and Mayer (2010) found no differences in learning outcomes between open-ended and menu-based self-explanation types.

To better understand the impact of each type of prompt, it is important to explore which learning behaviors, processes, and strategies that emerge during open-ended and focused self-explanations contribute to actively constructing knowledge. One factor impacting successful math learning and achievement is the student’s confidence or self-efficacy (Pan et al., 2022). However, students’ judgments of their abilities are vulnerable to cognitive biases, possibly skewing their perceptions.

Cognitive biases stem from heuristic strategies, or decisional short cuts (Blumenthal-Barby & Krieger, 2015), to reduce the amount of information considered when making a decision (Schwenk, 1986). The emergence of biases is mediated by contextual factors such as attributed risk (Halpern, 1989), age and expertise (Forbes, 2005), and time constraints (Lehner et al., 1997), and these biases can impact a student’s confidence, interest,
and achievement in math. For example, when students have high confidence but low knowledge (Pseudo-Confidence), they tend to experience a higher occurrence of confusion and frustration during learning (Di Leo et al., 2019).

Previous studies have also suggested that low-knowledge learners may attempt to alleviate confusion by avoiding activities that require deep thought (Rodrigo et al., 2007); for instance, gaming the system, where a student exploits a platform’s help/feedback systems to succeed instead of attempting to learn (Baker et al., 2006), and Without Thinking Fastidiously (WTF) behaviors, where a student engages in activities that are unrelated to the learning goal (Wixon et al., 2012).

In this paper, we extend previous findings by examining pseudo-confident refusal, a cognitive bias where a student refuses to self-explain and instead asserts their correctness despite system feedback indicating otherwise. Specifically, the current study seeks to identify and examine the role of pseudo-confidence during self-explanations on students’ affect, learning, and performance within a game-based learning environment (GBLE) called Decimal Point. Focused and open-ended prompts were used to elicit students’ problem-solving processes in math and identify learning behaviors that emerged during knowledge construction. Through the use of qualitatively coded text logs of student interactions, the current study investigated the influence of pseudo-confidence, on students’ affect, performance, self-efficacy, as well as interaction with the platform.

2. Methods

Decimal Point

The platform used in this study was Decimal Point, a GBLE that is based on an amusement park metaphor. Players were guided through different sets of mini games built within the Cognitive Tutor Authoring Tools (CTAT; Aleven et al., 2009) to learn about decimals. The students played a total of 24 mini games during regular class periods (across an average of 3-5 days) which lasted around 45 to 55 minutes each.

In the current study, players used the erroneous examples play mode, where they were required to find and fix common errors in given math problems and then self-explain their problem-solving process either with or without scaffolding (“Think about how you came up with this answer, and drag the correct option to the following blank(s).” versus “How do you know?”, respectively). The self-explanation prompts were designed to actively engage a student to develop a deeper understanding of critical concepts in mathematics. Each student was randomly assigned to one of three self-explanation conditions: menu-based, scaffolded, or focused format. In this paper, we study the focused, open-ended, self-explanations, which prompted students to explain their response with minimal guidance.

Sample and Student Level Metrics

Data were collected from a middle school in Pennsylvania between March and November 2021. Three hundred and fifty-eight ($n = 358$; 44% females) students participated in the study; however, only 85 were retained for the current set of analyses due to data completeness across variables of interest. Students participated in a survey before and after they played Decimal Point. The survey measured Decimal Efficacy, which indicates a student’s level of self-efficacy or their confidence in using decimal operations. The survey was adapted from literature (Pintrich et al. 1993) to better align with the context of the learning session, as used in Hou et al. (2020). Students were asked to rate statements on a 5-point Likert scale. We averaged three items to compute the Decimal Efficacy score: 1) I can do an excellent job on decimal number math assignments, 2) I can understand the most difficult material presented in decimal number lessons, and 3) I can master the skills being taught in decimal number lessons.

Students’ interaction data were examined using text replays (Baker et al., 2006), which display interaction logs from student platform usage in an easy-to-read form. Text replays
were then investigated to identify potential behaviors that emerged from cognitive biases and experiences of confusion (when a student is either confused or frustrated; Mogessie et al., 2020). This method has been used in previous studies to label a variety of student variables (Baker et al. 2006; DiCerbo and Kidwai, 2013; Zhang et al., 2022) quickly and with high inter-rater reliability. For this study, analyses were focused on self-explanation behaviors that occurred within the problem-solving process. As such, the log data were divided into clips at the level of entire problems.

**Construct Operationalization and Qualitative Coding**

The team first analyzed student responses for indicators of self-explanation behaviors using qualitative categories. This approach followed the recursive, iterative process used in Weston et al. (2001), including reviewing literature on previous work using the Decimal Point platform (Forlizzi et al., 2014) and student engagement behaviors within GBLEs (Zhang et al., 2022). Using grounded theory (Charmaz, 1983), we identified common behaviors that were indicative of 1) struggles involving self-explanation from the current mathematical content and 2) were indicative of difficulty generating accurate responses for the given problem.

Following the process used in (Zhang et al., 2022), two coders (1st and 4th authors) coded a set of 900 clips together, identified four behaviors occurring during self-explanations (see details in Table 1), outlined the criteria for each indicator, and created a rubric. The coding manual was reviewed and discussed by the research team to ensure a shared understanding of the criteria and constructs being studied. This process was repeated until the whole team had a mutual understanding of the codebook’s criteria and constructs.

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confusion</td>
<td>The learner explicitly mentions they do not know what to do or how to answer the problem and makes repeated answer submissions that are incorrect.</td>
</tr>
<tr>
<td>k = .68</td>
<td></td>
</tr>
<tr>
<td>Without Thinking Fastidiously (WTF) k = .89</td>
<td>The learner does not acknowledge the prompts for finding a solution and instead inputs text unrelated to the task at hand. (e.g., “F*** off,” “Stop asking me,” “Ohmygahd”)</td>
</tr>
<tr>
<td>Incongruent Response k = .62</td>
<td>The learner inputs correct information, but it is phrased differently from responses the system can recognize. These responses often involve minimal changes between response submissions where changes do not semantically change the answer; the student is trying to find a way to phrase the answer so that the system will accept it.</td>
</tr>
<tr>
<td>Pseudo-Confident Refusal k = .79</td>
<td>The learner refuses to explain their answers or mathematical process and instead refers to their inherent capacity to understand the mathematical problems by citing themselves as an authority to know the final answer. (e.g., “I know because it’s the right answer;” “Because I’m smart”)</td>
</tr>
</tbody>
</table>

### 3. Results

Spearman correlations with Benjamini-Hochberg post-hoc controls were computed to find associations between pseudo-confidence, performance metrics, and decimal efficacy survey responses. Students who demonstrated more pseudo-confidence in their self-explanations were more likely to submit incongruent responses ($r = 0.29$, adj. $\alpha = 0.01$, $p = 0.008$) and exhibit confusion ($r = 0.33$, adj. $\alpha = 0.007$, $p = 0.002$). In addition, pseudo-confidence in self-explanations were moderately and positively associated with making errors ($r = 0.25$, adj. $\alpha = 0.02$, $p = 0.02$) and taking longer to finish the game ($r = 0.37$, adj. $\alpha = 0.003$, $p = 0.004$). Students who demonstrated more pseudo-confidence in their self-explanations tended to score higher on Post-test Decimal Efficacy surveys ($r = 0.25$, adj. $\alpha = 0.025$, $p = 0.023$) and had increases in the normalized change in self-efficacy from before to after learning with the game ($r = 0.24$, adj. $\alpha = 0.028$, $p = 0.028$). These results implied that pseudo-confident responses may emerge in response to challenges responding to self-explanation prompts. Experiences of confusion were related to increased errors and time spent problem solving, and in response, students may turn to pseudo-confident responses to avoid further engaging
with the material. What is interesting to note is that pseudo-confident students were also more likely to positively estimate their decimal self-efficacy. This relationship may point to a tendency of pseudo-confident students to overestimate their abilities despite generating multiple errors and needing more time to solve math problems.

Table 2. Correlations for behaviors, achievement metrics, and survey level responses

A linear regression predicting the rate of pseudo-confident responses using the interaction between time spent problem-solving and pre-test scores of achievement revealed that increased pre-test scores positively predicted pseudo-confidence ($\beta = .02, p = .022$), suggesting that students who were more familiar with the content were more likely to assert their perceived correctness and refuse to explain. A significant interaction was also observed between pre-test and time: students with higher pre-test scores who spent a longer amount of time problem solving also applied more pseudo-confident responses ($\beta = .28, p = .018$).

Table 3 shows that across both achievement metrics (post-test and learning gains), a student who spent more time problem solving was related to decreases in their math performance. Students who gave more pseudo-confident responses were less likely to perform well on the post-test math assessment. A significant interaction was observed between time and the rate of pseudo-confident behaviors for predicting post-test. There is a complicated relationship here -- a simple slopes analysis revealed a significant relationship where pseudo-confident students who spent less time were less likely to have increases in their post-test scores. However, pseudo-confident students who spent more time did better.
4. Discussion

This study examined pseudo-confidence, a cognitive bias impacting math learning, and its relation to affect (e.g., confusion) and other learning behaviors (e.g., WTF, incongruent responses) during open-ended and focused self-explanations within a GBLE. We found positive associations between pseudo-confident responses with time spent problem solving and errors made during game-based learning about math. The positive correlation between these variables indicated a potential struggle self-explaining problem solving, where students may defer to heuristics to complete the activity. Additionally, students may demonstrate pseudo-confident responses as a means of self-presentation and impression management around math difficulties (Schwenk, 1986).

We also found that higher pseudo-confidence was associated with more confusion and incongruent responses during game-based learning. These relationships can be explained by the student failing to actively construct their knowledge when self-explaining, contributing to more cognitive incongruencies (i.e., confusion) or more incongruent responses (Di Leo et al., 2020). If a student made multiple errors during problem solving, but continued to refuse to explain their problem-solving approach (pseudo-confident), this may result in more confusion in the long term. The results also showed that the more pseudo-confident a student was, the higher their self-reported Decimal Efficacy was after game-based learning, indicating that responding in this fashion may increase student efficacy despite lower knowledge.

Next, we examined whether pre-test scores and time spent in the game impacted the rate of pseudo-confidence in self-explanations. The results showed that students with higher pre-test scores also had higher rates of pseudo-confidence in their self-explanations. This finding suggested that students who had more prior knowledge of math topics demonstrated more pseudo-confidence. Students with higher prior knowledge may be vulnerable to the Dunning-Kruger effect; they may be knowledgeable enough to be confident in their responses, but unable to detect blind spots in their knowledge.

Lastly, linear regression analyses revealed that the more time spent in the game and pseudo-confident responses made, the worse students performed on the post-test assessment. A simple slopes analysis revealed that pseudo-confident students who spend less time problem solving performed worse on the post-test assessment. These findings may align with previous research which suggest that students tend to make less effective decisions when required to solve problems or make decisions with less time (Lehner et al., 1997). However, pseudo-confident students who spent more time performed better on the post-test, possibly suggesting that this subset of students was still learning.

This study used a mixed-methods approach to investigate pseudo-confidence while students explained their approach to solving math problems during game-based learning. While self-report data outside the context of the activity was valuable for measuring global confidence in math, this approach presents challenges for studying the role of confidence during math learning. Confidence is a dynamic construct, one that changes during a learning session based on a student’s goals and in response to various types of feedback they may receive (e.g., time left or confusion; Pavlas, 2010). Thus, real-time measurement of confidence is important for advancing our understanding on if, when, and how confidence evolves during game-based learning about math. Future research should evaluate the role of other common cognitive biases that impact math learning, including examining the presence of stereotype threat or the hard-easy effect (overconfidence on harder problems) during math learning. For example, to what extent does the rate of pseudo-confidence change depending on how difficult the math problems are? This research has implications for identifying key learning behaviors and building adaptive self-explanation prompts and interventions that maximize students actively constructing knowledge about math with GBLEs.

References


GaMINLab - Meaningful gamification to engage students in science inquiry practices through simulation labs

Archana RANE\textsuperscript{ab*}, Sahana MURTHY\textsuperscript{a} & Sasikumar M\textsuperscript{b}
\textsuperscript{a}Indian Institute of Technology, Bombay, India
\textsuperscript{b}CDAC, Mumbai, India
*archana.rane@iitb.ac.in

Abstract: Practical labs in science are expected to foster disciplinary practices of “doing science” in learners. Simulation-based labs which provide the convenience of conducting practical lab experiments online however, come with their own set of challenges such as steep learning curves, learner working in isolation, lack of personalization which may adversely affect learner engagement and motivation. OLabs is a set of simulation-based labs for schools widely used in India. The main goal of this work is to redesign simulation-based labs like OLabs to engage learners in science practices and to motivate learners to engage in science practices. Meaningful gamification is one of approaches that focuses on intrinsically motivating learners to find meaning in the given learning context. GaMINLab, which is this proposed redesign of simulation-based labs like OLabs, was informed by literature recommendations for fostering disciplinary practices, addressing challenges in simulation-based labs, and meaningful gamification. We discuss the design features of GaMINLab.

Keywords: Simulation-based labs, meaningful gamification, science disciplinary practices, inquiry labs, scientific inquiry

1. Introduction

Practical labs, are expected to help learners learn science, learn about science and learn to do science. In recent years, the extensive emergence of simulation-based labs has facilitated learners to do virtual experiments anytime, anywhere, and any number of times. These labs also offer possibilities to illustrate concepts beyond the boundaries of traditional labs, for example illustrating magnetic field intensity, lifecycle of mosquito, etc. Despite world-wide proliferation and advantages, there are reported challenges such as learner isolation, steep learning curves, lack of suitable guidance, lack of personalization, having cookbook-based exercises, etc. The challenges may lead to frustration among learners, resulting in early dropout or loss of motivation and engagement. Thus, it is critical to give adequate attention to engage and motivate learners to do the underlying lab related activities.

OLabs is a significant initiative by the government of India to address concerns such as lack of basic lab infrastructure, limited practice opportunities, etc and supplement physical labs (M Sasikumar, 2016). OLabs, which is a set of simulation-based labs in various subjects at school level, has been adopted by thousands of state schools in India and lakhs of users access it on a daily basis. In its current form there are no tasks/problems built-in to engage learners in science disciplinary practices. The goal of this work is to re-design simulation-based labs like OLabs to a) provide opportunities for engaging in science disciplinary practices b) engage and motivate learners as they engage in these practices. This re-design, called GaMINLab (Gamification (Meaningful) in Inquiry Labs) is informed by literature recommendations for a) fostering disciplinary practices, b) addressing challenges in simulation-based labs, and theory of meaningful gamification. In this paper, we discuss GaMINLab design features.
2. Literature & Theory
Simulation-based labs facilitate conducting experiments online, overcoming the geographical and time limitations of traditional labs. To tackle challenges like steep learning curves, isolation, lack of personalization, design guidelines are recommended in literature a) providing investigation opportunities beyond the classroom, b) posing driving questions to focus learners' exploration, c) implicit scaffolding, d) encouraging reflection on findings, and e) promoting peer interaction and engaging in collaborative work (Moore et al., 2013).

Science education is increasingly focusing on involving learners in disciplinary practices, as emphasized in global curriculum recommendations. These practices include asking questions, developing and using models, planning and conducting investigations, analyzing and interpreting data, using mathematics and computational thinking, constructing explanations, engaging in evidence-based arguments, and obtaining, evaluating, and communicating information (Jaber et al., 2018). To foster disciplinary practices in a learning environment, literature suggests guidelines such as: a) providing direct experience with phenomena, b) making learner thinking visible to encourage grappling with ideas for constructing meaning, and c) assigning tasks with essential procedures and relevant resources for completing investigations. Inquiry learning is recommended for effective science learning, where students engage in sense-making, discussions, evidence-based explanations, etc. (Jaber et al., 2018) (Moore et al., 2013).

One of the goals for redesigning OLabs is to motivate learners to engage in lab related activities. Gamification, an established approach to fostering student engagement, uses game attributes to encourage game-like behavior in non-game contexts. Game-based mechanics can effectively engage learners, motivate their actions, and aid problem-solving. Reward-based gamification, a popular type of gamification, mainly relies on extrinsic motivation and may not yield lasting changes, and can be potentially unsatisfactory for some learners. Meaningful gamification, another type, employs game design elements to cultivate intrinsic motivation in non-game settings by creating an enjoyable learning environment where participants can explore and find meaning (Nicholson, 2015). This approach aims for sustained engagement compared to the short-term, reward-driven nature of extrinsic motivation. Meaningful gamification is based on self-determination theory (Ryan & Deci, 2000), which mentions competence, autonomy, and relatedness to form intrinsic motivation. The theory highlights that when these three psychological needs are fulfilled, people find tasks meaningful and continue participating.

3. Design
We propose a framework GaMINLab - Gamification (Meaningful) in INquiry Labs which derives its basis from the research recommendations for fostering disciplinary practices in science inquiry labs, theory of meaningful gamification. The design features are listed below:

- The learner will be given challenges in form of problems to be solved, each categorized in different problem scenarios like park, beach, etc. The system prompts learner for problem-solving activities that can be attempted in quest to solve the chosen problem.

- Learner is prompted for problem-solving activities which can be attempted in question to solve problem. These activities are adaptation of the problem-solving inquiry framework in technology enhanced classroom (Kim & Hannafin, 2011). Exploration, Reconstruction, Presentation/Communication are broadly mapped to tasks Investigate using the Lab, Propose Answers and predictions and Share and Discuss your Lab Investigation Report. The learner is provided access to specific OLabs simulation lab to investigate, make predictions, etc.in quest to solve the given problem. It is embedded in specific tasks.

- Learner is encouraged to prepare artifacts like investigation plan, interim and final lab investigation report comprising investigation plan, observations, proposed answers with reasoning, etc. Artifacts can be self, peer or teacher evaluated using given rubrics.

- There is no predefined sequence of problem-solving activities and at every stage the system prompts the learner with options to proceed further and learner can choose any one from those options, thus making a ‘choice’ in strategy to solve the problem. Potentially a learner can view the problem and directly jump to attempting the final assessment or go...
through one or more problem solving tasks or restart all over again. Motivating learners to use variants & try other aspects of lab is an important need.

- Each disciplinary task completion is acknowledged by system in form of corresponding badge. These acknowledgments given in recognition to them as they are progressing towards their final goal fulfills their needs of mastery and also motivates them to try further problems, and activities including variants.

- Lab onboarding activities make learner familiar with the UI objects in the lab, lab variants, actions needed for taking observations, activities include interactive content using hotspots, quiz and awards badge on completion.

- Chat option is given to discuss ideas, clarify misunderstanding regarding predictions, investigations with peers. They can also share lab investigation reports with peers.

- In the system scaffolds are provided when learners are preparing the investigation plan and lab investigation report. When learners choose the options from alternate paths, system provides brief information indicating what can be expected in the chosen stage which will help learner to move ahead or retrack the action.

Conclusion
GaMINLab, a re-design of OLabs, offers opportunities for learners to engage participate in disciplinary practices such as planning and executing investigations within the simulation lab, analyzing and interpreting recorded observations, constructing explanations based on evidence (observations), and communicating information by discussing and sharing the lab reports with peers. Collaborative discussions address learners' need for connection, preventing feelings of isolation during tasks. Making autonomous choices regarding which path to follow and which problem to solve fulfills their psychological desire to make self-determined choices, autonomy. The freedom to retry or restart reduces the fear of failure, motivating learners to persist in solving problems even after initial setbacks. Acknowledgment in form of badges motivates them to repeatedly engage in various problem-solving activities. Lab onboarding tasks and scaffolding help prevent early frustration during initial lab encounters. The inclusion of badges and lab onboarding supports the learners' pursuit of mastery. Therefore, the design of GaMINLab seeks to intrinsically motivate learners (by fulfilling their needs for autonomy, mastery, and relatedness) to actively engage in given scientific practices.

Currently GaMINLab is designed with two labs with one problem each. To validate the design a pilot study with 10-15 students is being planned. In this study we aim to gather information related to pattern of engagement of learners, motivation, degree of disciplinary practices reflected in learner artifacts, learner perception, usability etc. Study-2 will be conducted for larger set of learners (50-60), after incorporating suitable changes in system based on the findings of pilot study.

Acknowledgement
We would like to thank members from ETU, CDAC and IDP-ET department, IIT Bombay

References
A Skill Tracing Model for Player Character Control in STG

Peizhe HUANG*, Wanxiang LI, Wen GU, Kouichi OTA, Shinobu HASEGAWA

Division of Advanced Science and Technology, JAIST, Japan
Mobilus Corporation, Japan
Center for Innovative Distance Education and Research, JAIST, Japan
{s2320023, wgu, ota, hasegawa}@jaist.ac.jp, lwx9688@gmail.com

Abstract: STGs, a longstanding video game subgenre, have grown more intricate over time, deterring new players. To address this, a training system is required to improve character control skills in STG games. Bayesian Knowledge Tracing (BKT) is a common approach researchers use to monitor and assess students' progress. While BKT is effective in evaluating intellectual knowledge, it falls short in assessing character control skills in STG, a form of motion knowledge. This study proposes a Skill Tracing (ST) model that combines BKT approaches to monitor both cognitive knowledge and character control abilities. Results indicate its superiority in skill-tracking tasks over traditional BKT, offering a more accurate prediction of players' skill levels.

Keywords: STGs, Bayesian Knowledge Tracing, skill modeling, character movement classification

1. Introduction

STG, a challenging video game subgenre, demands exceptional reaction speed and skill mastery. Despite a loyal fanbase, its complexity hampers novice skill acquisition, affecting their experience. Matching challenge to skill level is vital for enjoyment (Sweetser & Wyeth, 2005), making skill training crucial. Traditional methods like repetitive practice are time-consuming and unspecific. Designing a focused training system for STG novices is essential for rapid skill mastery, requiring skill learning tracing.

Skill tracing constitutes a vital component of the STG skill training system, allowing it to comprehend a player's learning progress and skill proficiency. The Bayesian knowledge tracing (BKT) method (Corbett & Anderson, 1994) is commonly employed to model and monitor a student's mastery of knowledge. While in most concept-based learning scenarios, researchers merely assess whether a student has comprehended a concept or not, the situation differs when a player aims to master a gaming skill. In this context, it becomes important to not only gauge a player's familiarity with concepts such as game tips or tactics, but also to evaluate whether the player can effectively apply this information during gameplay, drawing upon their physical capabilities. This distinction implies that BKT might not be optimal for accurately modeling player skills.

This research aims to extend the BKT model to monitor player control skill learning in STG game. To achieve this, the following research questions will be addressed:

1. How can the BKT model be extended from knowledge learning tracing to encompass skill learning tracing?
2. Does the proposed model outperform the BKT model in the context of STG skill tracing?

2. Proposed Method

2.1 Skill Tracing Model
In skill learning, applying acquired knowledge involves the body's physical ability to become adept at the technique, while knowledge learning involves purely intellectual engagement. Physical ability here refers to coordinating bodily functions for motor tasks, encompassing coordination, reflexes, balance, and spatial orientation.

Assuming skill mastery ('S' in Equation (1)) results from the overlap of conceptual understanding ('K' in equation (1)) and physical familiarity ('B' in Equation (1)), it can be represented as: \[ S = K \cap B \] (1)

We proposed a Skill Tracing model (ST) based on the BKT model and Equation (1). Figure 1 is the structure of the ST model. The parameters comprise two sets of \([P(L), P(T), P(G), P(S)]\), each belonging to a distinct BKT model. These four parameters represent the probabilities of prior knowledge, learning new material, guessing, and making errors during the learning process. We combine the predictions of the two BKTs and determine the player's skill mastery state \(S\) using Equation (1). Two sets of responses, \(P_k\) and \(P_b\), are gathered from the observable layer. These responses are obtained from the same game stage at each step of the learning sequence. \(P_k\) denotes whether a player used a tactic in the game, helping predict the knowledge level (K) representing their understanding of that specific tactic in a stage. Tactic use is detected by analyzing keypresses and character movement. Treating a stage as a time series, the MrSQM time series classifier (Nguyen, T. L., & Ifrim, G., 2021) is employed to identify tactic application. \(P_b\) indicates whether a player completed a stage, and it's used to predict the body ability (B), indicating the player's capacity to finish a game stage at a specific difficulty level.

### 2.2 Game environment and skill setting

In this study, the gaming environment is the self-developed game "stgST," created using the open-source game engine LuaSTG (LuaSTG Wiki, n.d.). StgST comprises 18 stages with distinct bullet patterns outlined in Table 1. Our research centers on character control skills, vital for survival in the game. In this context, we define skills as employing specific tactics to dodge corresponding bullet patterns as shown in Table 1.

<table>
<thead>
<tr>
<th>Pattern name</th>
<th>Tactic</th>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aimed Shot</td>
<td>Tiny Moving</td>
<td>Use tiny moving to avoid aim-shot</td>
</tr>
<tr>
<td></td>
<td>Restream</td>
<td>Use restream movement to avoid aim-shot in screen edge</td>
</tr>
<tr>
<td>Round Shot</td>
<td>Circling</td>
<td>Use circling movement to avoid round shot</td>
</tr>
<tr>
<td>Bottom Bullet</td>
<td>Upper Avoiding</td>
<td>Move to upper screen to avoid bottom bullet</td>
</tr>
<tr>
<td>Revenge Bullet</td>
<td>Hold Fire</td>
<td>Stop shooting to avoid revenge bullet</td>
</tr>
</tbody>
</table>
3. Experiment and Result

We conducted an experiment to collect player data, recording a total of 957 game rounds from 21 players. The author manually recorded instances of player failures per round, while keypress information and character movement were automatically recorded. We assumed a stage to be cleared if a player failed in it fewer than four times. Player keypress and character movement were categorized into different patterns using a pre-trained MrSQM model, derived from author-generated play data.

The player data were modeled using the PyBKT Python library (Badrinath, A., Wang, F., & Pardos, Z., 2021). We compared the modeling outcomes of the ST and BKT models on player data, additionally assessing the impact of introducing the "forgets" parameter. The "forgets" parameter, introduced by Qiu, Y. et al. (2011), signifies the tendency for students to forget knowledge at a certain rate during the learning process.

Taking into account the imbalanced input data, we employed F1 score and balanced accuracy as metrics to assess the models. As indicated in Table 3, the results demonstrate that irrespective of the presence of the "forgets" parameter, the ST model surpasses the BKT model in F1 score by approximately 7% and in balanced accuracy by about 5.5%.

Table 2. F1 score and balanced accuracy of four models.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
<th>Balanced Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BKT</td>
<td>0.500</td>
<td>0.684</td>
</tr>
<tr>
<td>BKT + forgets</td>
<td>0.511</td>
<td>0.695</td>
</tr>
<tr>
<td>ST</td>
<td>0.570</td>
<td>0.747</td>
</tr>
<tr>
<td>ST + forgets</td>
<td>0.577</td>
<td>0.751</td>
</tr>
</tbody>
</table>

4. Conclusion & Future Work

In this study, we proposed a new model called the ST model to track a player's progress on the task of character control skill. The ST model is an extension of the BKT model, differentiating between knowledge-learning and skill-learning tasks. We validated the model on the player character control skill-learning task and found that it outperforms BKT. In addition, we found that the parameter forget had essentially no effect on the results of the experiment, probably because the duration of the experiment was so short that players were unlikely to forget.

In future research, we would like to complete a skill training system for novice STG players. Additionally, we plan to explore the potential application of the ST model to other types of skill-learning tasks beyond game-based scenarios.

References

The Design and Practice of Scientific Inquiry Activities for Children Aged 5-6 Based on an AR Flashcard Environment

Tian-Tian GOU, Min-Sheng FAN* & Bin-Li WANG
School of Educational Technology, Northwest Normal University, China
*xmanfms@126.com

Abstract: In this study, a design framework for scientific inquiry activities for children aged 5-6 based on an AR flashcard environment is constructed, cases of scientific inquiry activities are designed according to this framework, and quasi-experimental studies are carried out. And it selects two large classes of students in a kindergarten in Xiahe County, Gansu Province of China as experimental subjects, and sets up an experimental class and a control class, with 20 students each. The research results indicate that compared with the traditional learning environment, the implementation effect of 5-6 children's scientific inquiry activities based on an AR flashcard environment is obvious. AR technology fully mobilizes children's interest and enthusiasm for inquiry, obtain a better inquiry experience, and thus improve children's scientific inquiry abilities.

Keywords: AR flashcard, children 's science education,, children 's scientific inquiry abilities

1. Introduction

In April 2022, the Ministry of Education issued the "Science Curriculum Standards for Primary Schools of Compulsory Education". Subsequently, the Ministry of Education of China issued the "Guidelines for the Learning and Development of Children Aged 3-6", and the Guide emphasizes that the essence of children's science education is to stimulate children's interest in inquiry, allowing them to experience the process of inquiry, and cultivating their early inquiry abilities (Zhanlan Liu, 2008). Children's scientific inquiry activities belong to the important content of kindergarten science education. Generally speaking, children's scientific inquiry activities are divided into formal and informal scientific inquiry activities (Lan Lu, & Mei Hang, 2011).

However, there are many urgent problems that need to be solved in children's scientific inquiry activities in reality, such as the "formalization" of inquiry, the lack of necessary learning resources, which are not conducive to the development of children's scientific inquiry abilities. In the virtual and real environment created by AR flashcards, children can interact with the 3D models in real-time after scanning the AR flash card through the smart mobile device, which can stimulate children's interest in learning and inquiry, and put themselves in a safe environment for experimental operation and independent inquiry, thereby improving the problems existing in children's science activities.

Based on the above reasons, this study aims to apply the AR flash card-based learning environment to the design and implementation of scientific inquiry activities for children aged 5-6, mainly studying the following issues:

1. How to design a framework for scientific inquiry activities for children aged 5-6 based on the AR flashcard environment?
2. Can the framework for scientific inquiry activities for children aged 5-6 based on the AR flashcard environment effectively improve children's scientific inquiry abilities?
2. Literature Review

2.1 AR Technology Features and its application in early childhood education

AR technology can interact with virtual and real objects due to its characteristics, with the technical characteristics of AR, the use of technology tools in education can provide new learning opportunities to increase the interaction between individuals and entertainment learning, bring a rich learning experience and make the learning process more positive, effective and meaningful (Alsumaiti, & and Musawi, 2013).

Through literature review, it is found that the current expression of early childhood education content is mainly cognitive cards, games, picture books, etc. AR technology blends the physical and virtual worlds to provide a new learning platform, and the technology is now widely used in the educational environment of elementary, high school and college students. In the study, the experimental results showed that AR improved students' learning motivation, engagement and achievement compared with traditional teaching media (Bacca, J. et al., 2014). Similarly, in one survey, students were found to understand and remember content and perform better, cooperate more, and show more motivation and focus when using AR compared to non-AR traditions or media (Radu I, 2014). According to the above statement due to the problems encountered by children aged 5-6 in the implementation of scientific activities, as well as the current goals and requirements of children's science education and the application of AR technology, AR technology is undoubtedly a suitable tool or product for children's learning.

3. Design of Scientific Inquiry Activities for Children Aged 5-6 Based on an AR Flashcard Environment

3.1 Framework Construction

This study aims to improve children's scientific inquiry abilities, and according to the ADDIE teaching design model (Merrienboer, 1997), from the five stages of analysis, design, development, implementation, and evaluation, this study benchmarks the preliminary analysis, activity process design, activity resource development and preparation, and activity case implementation of children's scientific inquiry activities based on an AR flashcard environment. A design framework in Figure 1 for scientific inquiry activities for children aged 5-6 based on an AR flashcard environment is constructed. Among them, in the activity process design stage, the "5E" teaching modethat is in line with the goals and concepts of science education is selected as the basis, which is a new teaching method developed by American educators in the field of biology (BSCS), which is five links: Engagement, Exploration, Explanation, Elaboration, and Evaluation.

![Figure 1. A Framework for Scientific Inquiry Activities for Children Aged 5-6 on an AR Flashcard Environment](image-url)
### 3.2 Learning Activities Design

#### 3.2.1 Activity Case Design

The theme of this research activity is selected from the "Harmonious development curriculum teacher's Book". Table 1 is a concrete example of an activity case.

<table>
<thead>
<tr>
<th>Table 1. Activity Case Design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity theme</strong></td>
</tr>
<tr>
<td><strong>Context import</strong> (Create problem situations to stimulate children's interest)</td>
</tr>
<tr>
<td><strong>Ask questions Make assumptions</strong> (Throw questions, elicit activities)</td>
</tr>
<tr>
<td><strong>Explore operations</strong> (Children use AR tools to explore how animals survive the winter)</td>
</tr>
<tr>
<td><strong>Share and communicate</strong> (Share findings)</td>
</tr>
<tr>
<td><strong>Explain doubts</strong> (Teachers answer questions and integrate knowledge)</td>
</tr>
<tr>
<td><strong>Extended migration</strong> (Guide children to solve problems with new knowledge)</td>
</tr>
<tr>
<td><strong>Multidimensional evaluation</strong> (Self-evaluation and mutual evaluation are encouraged)</td>
</tr>
</tbody>
</table>
4. Methods

4.1 Research Methods and Design

This study adopts quasi-experiment and collects relevant data of children through scale and interview content before and after the experiment. Comprehensive qualitative and quantitative methods are used to analyze and compare whether the traditional learning environment and the AR flash card-based learning environment cause differences in the cultivation of children's scientific inquiry abilities, and whether the children's scientific inquiry abilities are significantly improved.

![Figure 3. Specific Experimental Procedures](image)

4.2 Research Objects and Implementation

In this study, two large groups in a kindergarten in Xia he County, Gansu Province are selected as experimental objects. The number of students in both the experimental and the control group is 20. The experimental group adopts the scientific inquiry activity process based on an AR flashcard environment to teach, while the control group adopts the traditional teaching environment, and conducts 30 teaching activities in total for 10 weeks. Before the experiment begins, the research tools are used to pre-test the scientific inquiry abilities of children, so as to ensure that the level of students in the two groups is similar and homogeneous. In the experimental group, two students cooperate to control the tablet and complete the exploration task. According to the teacher's explanation and demonstration as well as the student's performance, the post-test data is collected and analyzed after the teaching activities.

![Figure 4. Teaching Implementation Examples](image)

4.3 Measurement Tools

4.3.1 Test Scale

The test scale of this study is the "Evaluation Table of Children's Scientific Inquiry Abilities". Its observation indicators are divided into six dimensions. There are 5 grading scales under each
dimension. Its Chronbach’s Alpha coefficient is 0.814, indicating good reliability. The KMO value is 0.727, indicating that the validity meets the requirements.

4.3.2 Interview Outline

This study designs the teacher and children interview outlines. The teacher interview designs six questions, such as: “To implement scientific inquiry activities based on an AR flashcard environment, what are the changes in your teaching methods?” The children’s interview also designs six questions, such as: “What do the children learn in class today?”.

4.4 Data Analysis

4.4.1 Pre-experimental Homogeneous Analysis

The data in Table 2 shows that the Sig of the both of groups are all greater than 0.05 in six dimensions, which do not reach a significant difference. It shows that the experimental group and control group can carry out the next intervention experiment with the same level of scientific inquiry abilities.

Table 2. Homogeneous Analysis of Scientific Inquiry Abilities in Experimental and Control Groups

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>T</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curiosity and interest</td>
<td>Experimental group</td>
<td>20</td>
<td>10.50</td>
<td>3.591</td>
<td>.929</td>
<td>.359</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>9.50</td>
<td>3.204</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propensity for inquiry</td>
<td>Experimental group</td>
<td>20</td>
<td>10.25</td>
<td>3.024</td>
<td>.523</td>
<td>.604</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>9.75</td>
<td>3.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operability of materials</td>
<td>Experimental group</td>
<td>20</td>
<td>10.00</td>
<td>3.627</td>
<td>.213</td>
<td>.833</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>9.75</td>
<td>3.796</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth of inquiry</td>
<td>Experimental group</td>
<td>20</td>
<td>12.75</td>
<td>3.024</td>
<td>.284</td>
<td>.778</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>13.00</td>
<td>2.513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to record information</td>
<td>Experimental group</td>
<td>20</td>
<td>10.25</td>
<td>3.796</td>
<td>.237</td>
<td>.814</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>10.00</td>
<td>2.810</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to collaborate and communicate</td>
<td>Experimental group</td>
<td>20</td>
<td>11.00</td>
<td>3.078</td>
<td>.737</td>
<td>.466</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>11.75</td>
<td>3.354</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>Experimental group</td>
<td>20</td>
<td>10.50</td>
<td>3.591</td>
<td>.377</td>
<td>.708</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>9.50</td>
<td>3.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4.2 Post-experimental Significance of Differences Analysis

According to the data analysis results of Table 3, The average score of the experimental group is significantly higher than that of the control group, indicating that the implementation effect of the experimental group is better than that of the control group.

The results in Table 3 below of the independent samples T-test analysis of the experimental group and the control group shows that the significance of the T-test is 0.000<0.05, indicating that there is a significant difference between the average value of the total score of scientific inquiry abilities of students in the experimental group and the control group.

Table 3. Independent Samples T-test of Scientific Inquiry Abilities in Experimental and Control Groups

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>N</th>
<th>M (Relative score)</th>
<th>SD</th>
<th>F</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td>Experimental group</td>
<td>20</td>
<td>99.75</td>
<td>6.172</td>
<td>.854</td>
<td>3.531</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Control group</td>
<td>20</td>
<td>88.75</td>
<td>4.833</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
According to the T-test analysis results of paired samples in Table 4 below, the average score difference of six dimensions of scientific inquiry abilities is 28.75, and the correlation significance of the total score of paired samples is less than 0.05, indicating that there is a significant difference between children's total score of scientific inquiry abilities before and after the practice of inquiry activities.

Table 4. Paired Samples T-test of Scientific Inquiry Abilities in Experimental Group

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>M (Relative score)</th>
<th>df</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest-Total score</td>
<td>20</td>
<td>64.75</td>
<td>19</td>
<td>-1</td>
<td>0.000</td>
</tr>
<tr>
<td>Posttest-Total score</td>
<td>20</td>
<td>93.50</td>
<td></td>
<td>8.243</td>
<td></td>
</tr>
</tbody>
</table>

4.4.3 Interview Analysis

Through the analysis of interviews between teachers and young children, it can be seen that teachers are satisfied with the design and implementation of scientific inquiry activities based on the AR environment, children's scientific inquiry abilities have been improved in the experiment, and teachers are also willing to use AR in future activities. It was agreed that AR tools were more fun and convenient than previous activities and that children could explore and record in an active atmosphere.

5. Conclusion

Through this experiment, this study mainly draws the following conclusions:

1. The design framework of scientific inquiry activities for children aged 5-6 based on the AR flashcard is effective.
2. AR flashcards can promote the effective development of children's scientific inquiry activities.

References


Toward Game-Based Learning of Japanese Writing for Elementary School Students

Kazumasa OMURA\textsuperscript{a}, Kei KUBO\textsuperscript{b}, Frederic BERGERON\textsuperscript{c} & Sadao KUROHASHI\textsuperscript{a,d}

\textsuperscript{a}Graduate School of Informatics, Kyoto University, Japan
\textsuperscript{b}Center for Japanese Language, Waseda University, Japan
\textsuperscript{c}SystemI Co., Ltd., Japan \textsuperscript{d}National Institute of Informatics, Japan
\textsuperscript{*}omura@nlp.ist.i.kyoto-u.ac.jp

Abstract: It is a long-standing problem that many elementary school students in Japan have an aversion to writing compositions. To address the problem, we designed an AI educational game for elementary school students to study Japanese writing utilizing existing language resources. In the game, players construct simple and complex sentences by connecting given word cards with particle marks. The constructed sentences are automatically scored using large-scale language resources, allowing players to receive on-the-spot feedback, such as how to improve the use of a case marker. We also developed smartphone and web applications of the game and conducted a user study to assess it. The results of the user study demonstrated that our application can be used as a good introduction to studying Japanese writing.

Keywords: Game-based learning, Japanese writing, Language resources

1. Introduction

Since we use language as a means of cognition and communication, language education plays an essential role in our lives. In language education, written language production is also indispensable for learning how to express our thoughts.

However, in Japanese writing education, it is a long-standing problem that many elementary school students have an aversion to writing compositions (National Institute for Educational Policy Research, 2008; Ritsumeikan University Library, 2017). One of the possible reasons is that they learn Japanese writing only through open-ended writing assignments such as book reports and essays on topics related to daily life. Such assignments rarely motivate the students to go further than merely complete them, and they often struggle to come up with what to write due to the open-ended format. In addition, they receive little feedback on their writing, leading to a vicious cycle where they become increasingly aware of being poor at writing without learning how to improve.

In order to ameliorate the current situation, Japanese education is in need of an educational material that offers a more engaging experience, i.e., allows students to construct sentences with fun and receive feedback on their writing. However, there are two major challenges to achieving this: how to reduce an aversion to constructing sentences and how to evaluate writing automatically.

Game-based learning (Vandercruysse et al., 2012), which aims to make the learning process more fun with a game, is a promising solution to the first challenge. There has been no AI educational game for studying Japanese writing to the best of our knowledge; thus, it is worth attempting to develop such a game and investigate its effectiveness.

Regarding the second challenge, we focus on existing language resources. In natural language processing, large-scale language resources have been built so far to teach linguistic and world knowledge to computers (Fellbaum, 1998; Kawahara & Kurohashi, 2006; Speer et al., 2017), some of which can be utilized for human learning as well as machine learning. By exploiting them, it is now possible to automatically generate and score simple writing practice questions, and the aforementioned educational material has become more feasible.
Against the above background, we design an AI educational game for elementary school students to study Japanese writing utilizing existing language resources. Hereafter, we call it "Kotoba-musubi". ("Kotoba" and "musubi" are the Japanese words that mean "words" and "connecting", respectively.) Section 2 describes the details.

We also develop smartphone and web applications of Kotoba-musubi and conduct a user study to investigate its effectiveness. Section 3 presents the positive results showing the effectiveness in reducing an aversion to writing compositions. We have released the game for educational purposes, the link to which is provided in Section 4 with a conclusion.

2. Kotoba-musubi

2.1 Game Design

Kotoba-musubi is a word-based game, where the player constructs simple sentences by connecting content words with case markers and combining the simple sentences with discourse markers into complex sentences with contingent relations (cf. Figure 1). In order to construct sentences, players arrange given 12 word cards and nine particle marks according to the empty rectangle and oval frames, respectively.

![Figure 1. Screenshot of Kotoba-musubi with gloss (left) and its scoring results (right). “NOM” and “ACC” in the left figure stand for “nominative” and “accusative” cases, respectively.](image)

**Word Card:** The 12 word cards break down into six noun cards, five verb/adjective cards, and one wildcard that allows players to enter a word freely. The role of each card is distinguishable by its color or shape; thus, players can learn how to construct sentences even if they do not understand the concept of parts of speech. In addition, ruby characters are written above Chinese characters to facilitate reading. Regarding verb/adjective cards, the predicates can be conjugated in the present, past, negative, or past-negative form, which makes it possible to create more diverse sentences.

**Particle Mark:** The nine particle marks break down into five of major case particles (nominative, accusative, dative, and instrumental cases, and “with” or “and”) and four of discourse connectives representing contingent (causal and conditional) relations. As with the word cards, the role of each mark is distinguishable by its color or shape. Each mark is arrow-shaped, its direction indicating the dependency between words.

The score of each sentence is automatically computed by pattern matching with language resources, allowing players to receive instant feedback. For instance, if no examples of a simple sentence are found in the Japanese case frames (Kawahara & Kurohashi, 2006) due to incorrect usage of a case marker, our system gives feedback suggesting a more appropriate case marker (cf. Figure 1). The player’s objective is to achieve a higher score by constructing basic simple sentences and complex sentences with contingent relations.

2.2 Method for Building Word Card Sets

656
In order for the game to be more fun, it is preferable that players can construct several simple and complex sentences from a given word card set. To guarantee this, we focus on core events (Omura et al., 2020; Omura & Kurohashi, 2022).

Core events are defined as the high-frequency predicate-argument structures acquired from case frames (Kawahara & Kurohashi, 2006); that is, each core event consists of one predicate and a few of its frequently co-occurring arguments and case markers. We adopt core event pairs with contingent relations such as “雨が降る→長靴を履く (it rains→wear rain boots)”, which can be extracted at scale from the Kyoto University Commonsense Inference dataset (cf. https://nlp.ist.i.kyoto-u.ac.jp/EN/?KUCI).

Our method is to automatically generate word card sets from core event pairs. It roughly consists of the following three steps (cf. Figure 2).

**Step 1: Filter Core Event Pairs by Word Difficulty.** First, we adjust vocabulary using the Japanese word difficulty database (Mizutani et al., 2019), considering the target users are elementary school students. This database contains 26k words of basic Japanese vocabulary labeled with their average acquisition time using crowdsourcing. Acquisition time is regarded as word difficulty and expressed as a number from 1 to 5 (cf. Table 1).

<table>
<thead>
<tr>
<th>Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquisition time</td>
<td>Before ES</td>
<td>Early ES</td>
<td>Late ES</td>
<td>After JHS</td>
<td>Never seen or heard</td>
</tr>
</tbody>
</table>

In this study, we set the following threshold conditions.

- **Easy** (for early elementary school students): Maximum word difficulty of words in a core event pair does not exceed 2.0.
- **Medium** (for middle elementary school students): Average word difficulty of words in a core event pair exceeds 1.5, and maximum word difficulty does not exceed 2.5.
- **Hard** (for late elementary school students): Average word difficulty of words in a core event pair exceeds 2.0.

We exclude core event pairs not satisfying a threshold condition. We set a strict upper bound but allow easy words to get mixed in harder word card sets to smooth the change in difficulty.

**Step 2: Select Core Event Pairs.** Second, we select five core event pairs from the ones satisfying a threshold condition. Specifically, we take one core event pair (hereafter, seed) and randomly select the other four pairs that share the former or latter core event of the seed. If we fail to get five pairs, including a seed, we skip generating a word card set.

**Step 3: Generate a Word Card Set.** Lastly, we generate a word card set by breaking down the five pairs selected in the previous step into predicates and arguments. If we obtain six nouns and five or six verbs/adjectives after de-duplication, we regard the words as a word
card set. In case that six verbs/adjectives are obtained, we replace the least frequent verb/adjective with a wildcard.

2.3 Careful Examination of Core Event Pairs and Building Word Card Sets

We must be extremely careful not to give inappropriate questions as an educational application. Accordingly, we request two linguistic experts to manually classify the core event pairs used for word card sets into the following categories.

- **Valid**: A core event pair has contingent relation.
- **Invalid**: A core event pair has no contingent relation.
- **Inappropriate**: A core event pair contains educationally inappropriate expressions.

We use the “valid” core event pairs for generating word card sets.

Prior to the examination, we excluded core event pairs containing minor case particles or words unregistered in the Japanese word difficulty database. As a result, 9k core event pairs were left and examined. We also automatically assigned reading to each word and conjugated forms to each predicate using the Japanese morphological analyzer, Juman++ (Morita et al., 2015; Tolmachev et al., 2018). We asked the experts to correct auto-assigned readings in addition to the examination.

As a result, we obtained 4,362 valid, 3,560 invalid, and 1,120 inappropriate core event pairs; then, we generated word card sets from the verified core event pairs according to the method described in Section 2.2. Finally, we obtained 99 easy, 594 medium, and 284 hard word card sets, achieving a size that is sufficient for solo play.

2.4 Automatic Scoring

When players tap a scoring button, the information on how cards and marks are placed is sent to a back-end server and automatically scored. The automatic scoring is performed in the following three steps.

**Step 1: Recognize Simple and Complex Sentences.** A simple sentence is recognized as a sequence of a verb card at the end and noun cards in the rest connected by case particle marks, and a complex sentence as two simple sentences connected by a discourse connective mark. We consider all possible combinations of simple sentences.

**Step 2: Score Each Sentence Automatically.** Our policy is to prioritize simple sentences whose predicate and argument frequently co-occur (i.e., idiomatic) or whose length is longer. The score of each simple sentence is determined based on the number of examples in the Japanese case frames (Kawahara & Kurohashi, 2006). Specifically, regarding the case frame \( cf \) where examples are found, we compute a score of each argument and case pair \((a, c)\) with the following function \( S \) and sum it up.

\[
S(a, c, cf) = 0.5 + 0.5 \times \min (1, \frac{f_{a,c,cf}}{f_{cf}} \times 5)
\]

where \( f_{a,c,cf} \) is the frequency of argument \( a \) in case \( c \) of case frame \( cf \), and \( f_{cf} \) is the frequency of case frame \( cf \).

Each score is converted into a symbol for further interpretability. Specifically, a simple sentence is graded as “☆☆☆” if the score is 1 or more, as “☆” if the score is between 0 and 1, and as “?” if the score is 0. Regarding complex sentences, they are graded as “☆☆☆☆☆” if they contain one of the verified core event pairs; otherwise, as “?”.

When examining the containment relation, we take into account the polarity of negation of a predicate.

**Step 3: Generate Feedback based on the Results.** The following feedback is generated from the sentences graded as “?” depending on the reasons behind them.

- There is a more appropriate case marker that allows for the sentence to achieve a score.
- The sentence is grammatically incorrect (e.g., the sentence starts with a verb/adjective, a verb/adjective depends on a noun, and so on.)
- The sentence is not found in language resources.

Regarding the third point, we expect to collect unknown contingent relations through an error reporting function and thus improve our evaluation system.
3. User Study

We developed iOS, Android, and web applications of Kotoba-musubi and conducted a user study to assess engagement.

3.1 Settings

We recruited 80 pairs of elementary school students and their parents across Japan and had the children play the game for an hour in total over two days. 80 students consist of 10 boys and 10 girls each from third to sixth grade in elementary school. In order to let them play freely, we neither set their quota nor specify the difficulty of questions they tackled. After playing the game, they answered the questionnaire described in Table 2.

Table 2. Main items and options of the questionnaire answered by the participating children. Regarding the second and subsequent items, the option with a larger number is preferable.

<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you want to continue to play Kotoba-musubi?</td>
<td>5. Yes 4. Yes, a little 3. Neither 2. No, a little 1. No</td>
</tr>
<tr>
<td>Do you think it was worth playing Kotoba-musubi?</td>
<td>5. Yes 4. Yes, a little 3. Neither 2. No, a little 1. No</td>
</tr>
</tbody>
</table>

3.2 Results and Discussion

Figure 3 shows the aggregate results of the questionnaire. We can see that elementary school students tend to dislike writing compositions, as mentioned in Section 1. Despite this disadvantageous situation, 70% of the participating children enjoyed Kotoba-musubi, and 90% answered it was worth playing the game. In addition, 70% expressed their will to continue to play the game, which suggests that it is a good introduction to studying Japanese writing.

Figure 3. Aggregate results of the questionnaire described in Table 2. The numbers in the figure represent those of children who chose the option.

We also investigated the number of children who disliked writing compositions but enjoyed the game. We found that 18 of 37 children who disliked writing compositions enjoyed the game. The result supports effectiveness in reducing an aversion to writing compositions.

The excerpted feedback comments from the participating children are listed below.

- + It was fun and refreshing to create sentences using words I didn’t usually use.
- + I’d honestly like to see the game introduced to school tablets.
- ± It will be more fun if we can compete with our friends for higher scores.
- − I couldn’t fully understand how to play the game.
- − It was hard to get sentences scored, which made me frustrated.

While many children noted they enjoyed themselves, there were also several comments that it took time to understand how to play, which raises the need for improving a tutorial.
The user study also revealed some current issues of Kotoba-musubi. For instance, there is room for improvement in reducing the feeling of studying while playing the game. As one of the feedback comments suggests, we need to enhance the enjoyment of our application by introducing game elements such as a match game system. Another issue is the quality of automatic scoring. One of the possible remedies is to incorporate neural language models into automatic scoring for more flexible evaluation. We will explore how to utilize neural language models considering the computational cost.

4. Conclusion

We proposed an AI educational game for elementary school students to study Japanese writing, which fully utilizes existing language resources such as case frames, a word difficulty database, and a commonsense inference dataset. We also developed smartphone and web applications of the game and carried out a user study involving 80 pairs of elementary school students and their parents to assess it. The results of the user study demonstrated the effectiveness in reducing an aversion to writing compositions. We expect elementary school students to develop their vocabulary and reasoning skills in a ludic manner with the game. The link to the game and the supplementary material of this paper are available at https://nlp.ist.i.kyoto-u.ac.jp/EN/?Kotobamusubi.

We will address the remaining issues referring to the feedback comments and further investigate long-term educational effects such as vocabulary development. We also consider collecting unknown contingent relations through error reports.

Acknowledgements

This work was supported by the Japan Kanji Aptitude Testing Foundation. The first author was also supported by Information/AI/Data Science Doctoral Fellowship of Kyoto University and Grant-in-Aid for JSPS Fellows #22J15958.

References

The Impact of Digital Game-based Learning with a Mathematical Game Application on Calculation Abilities of Grade 4 Students

Yin-Bei LIU*, Alex Wing Cheung TSEb*
Faculty of Education, The University of Hong Kong, Hong Kong
a*u3598295@connect.hku.hk; b*awctse@hku.hk

Abstract: Calculation abilities are essential in elementary school mathematics and learning with game applications have been demonstrated to improve student's learning of mathematics, which is conducive to developing their calculation abilities. As one of the core mathematical skills, calculation abilities can be improved through continuous calculation practice. However, there is currently just a little research on the effects of using digital game-based learning (DGBL) with a mathematical game application in elementary school for developing students’ calculation abilities. Therefore, 78 students participated in this quasi-experimental study to evaluate the possible influence of DGBL with a mathematical game application “Oral Math Hero” through iPads on the calculating abilities of fourth-grade students in a mainstream school in Mainland China. A mathematical game application was integrated into a four-week classroom activity for the experimental classes, with pre-test and post-test utilizing a standardized calculation abilities test: Abilita diCalcoloz Calculation Abilities-Memory and Training Group 6-11 (Cornoldi et al., 2002) for both experimental and control groups. Data analysis results using ANOVA showed that there were significant differences in students’ calculation abilities when learning with the mathematical game application on iPad in mathematical classrooms, with significant differences in overall calculation abilities between the experimental group (n=40) and the control group (n=38) in grade 4. In other words, we found that after practicing calculations with the mathematical game application, students were more likely to achieve better calculation abilities, especially in terms of faster calculation speed, and lower error rates. However, there were no significant differences in numerical knowledge, learning with this mathematical game application might not lead to acquiring more mathematical knowledge. This study provides a realistic perspective for elementary mathematics educators and teachers to understand the potential of learning with the mathematical game application: it can be an effective tool to improve the calculation abilities of students in grade 4. The second stage of this project is to explore the reasons behind the results of the study, revealing possible factors of DGBL with the mathematical game application that may facilitate certain aspects of calculation abilities. Further suggestions are made to integrate DGBL into the elementary mathematics classroom.

Keywords: Digital game-based learning, calculation abilities, mathematical games

1. Introduction

Students’ calculation refers to the abilities to understand patterns and relative quantities between numbers and to perform operations between numbers (addition, subtraction, multiplication, and division) in a more flexible way in mathematics (Feigenson et al., 2004; Tall & Dehaene, 1998). They are intrinsically important for mathematical performance at the elementary level (Cowan et al., 2011). Similar to elementary school mathematics curricula in different leading countries, grade 4 students have to master four operations (addition, subtraction, multiplication, and division) according to the latest curriculum standard in Mainland China, and they need to keep practicing calculation abilities to find easier solutions (Ministry of Education of the People’s Republic of China, 2022). Students' performance and
efficiency in calculation can be improved through constantly repeated practice (Imbo & Vandierendonck, 2008). However, repeated paper and pencil exercises are generally boring for students, and too many repetitions can make them lose interest to continue the practice and thus not effective to improve their calculation abilities (Chen et al., 2020). It can result in poorer mathematical skills and fewer arithmetic scores, which can have an impact on students’ future performance in mathematics (Núñez Castellar et al., 2014).

Digital Game-based Learning (DGBL) is a learning approach that uses digital games as a means of delivering educational material. The use of mobile game applications in the class can help students to participate more actively in educational activities and achieve higher instructional goals and outcomes (Vasiliou & Economides, 2007; García-Bárcena & García-Crespo, 2007) because of its funny nature. Students are more likely attracted to participate in and thus learn and acquire new knowledge and skills during the games (Corti, 2006). As past research has found that using DGBL in mathematics classes encouraged students to stay on track and increased their engagement with the learning process (Jabbar & Felicia, 2015; Tsai et al., 2012). However, it is not known whether the use of these game applications can improve students’ calculation abilities in mathematics. Besides, prolonged use of digital games for study may lead to game addiction, which can generate adverse effects on students’ academic performance and affect their physical and mental health (Eyimaya & Irmak, 2021). Thus, many past research on calculation abilities focused on cognitive factors and the relationship between previous academic skills and interests (e.g., Maedamuerk et al., 2018; Commodari & Di Blasi, 2014), or focused on kindergarten students’ numeracy (Tajudin et al., 2015; Zhu & Chiu, 2019), but less on the effects of digital games.

Therefore, the following research question is posed in this study: Does learning with a mathematical digital game affect the calculation abilities of grade 4 students?

2. Literature Review

2.1 Research gap derived from literature review

Research has found that digital game-based learning can be a promising educational approach that can help students easier to achieve targeted learning outcomes (Wang et al., 2022) and motivate students to learn because it generally provides timely feedback on task evaluation (Mathrani et al., 2016), promotes students’ interest in learning and increases classroom participation (Ke, 2008; Tsai et al., 2012). However, other research on DGBL are still inconclusive in this issue. For instance, there is no significant evidence that DGBL can improve students’ understanding of mathematical skills (Ke, 2008). Or Young et al. (2012) stated that limited data is supporting the validity of digital games in the field of mathematics.

According to recent studies, calculation abilities are important for students’ mathematics learning and even affect their later mathematics performance in secondary and post-secondary schools (e.g., Cahoon et al., 2021). However, previous studies on calculation abilities more tended to focus on factors such as attention, including selective visual attention and the ability to suppress distracting information, and how they affect the accuracy of verbal calculations in second and third-grade students (e.g., Commodari & Di Blasi, 2014; Bigozzi et al., 2021). Other studies have explored the impact of play environments on elementary school students’ flexible number knowledge and mathematical fluency, as well as the relationship between calculation and previous skills and interests (Brezovszky et al., 2019; Maedamuerk et al., 2018). There are only little studies on the effects of digital games on students’ calculation abilities at the elementary level though there are studies about the speed and accuracy of calculation for first and second-grade students (Núñez Castellar et al., 2014; van der Ven et al., 2017; Núñez Castellar et al., 2015).

To sum up, based on the literature review, there is still a research gap: few studies have examined the effects of digital game-based learning on fourth graders’ calculation abilities.
2.2 Theoretical framework

McCloskey et al. (1985) made a distinction between calculation abilities and the ability to comprehend as well as that of generating numbers. According to their model, the calculation system involved processing operational symbols, retrieving basic arithmetic facts, and performing calculated operations, in addition to numerical processing. These elements involved varying degrees in both mental and written calculations. Furthermore, any calculation work requires a certain level of number comprehension and generation (Commodari & Di Blasi, 2014). Based on Computing System Model, calculation abilities can be divided into three main dimensions: numerical knowledge, calculation accuracy, and calculation speed (Cornoldi & Lucangeli, 2004).

According to Dehaene et al. (1990), numerical knowledge is the set of abilities and knowledge that enables a student to comprehend numerical quantities and their conversions. A student’s ability to perform mathematical calculations quickly and accurately can be defined and determined by the concept of calculation speed and calculation accuracy respectively (Jordan & Montani, 1997). Based on these three key factors, Cornoldi et al. (2002) designed a standardized calculation abilities test: Abilita diCalcoloz Calculation Abilities- Memory and Training Group 6-11 (AC MT 6-11).

The strength of this framework is its ability to take the abstract concept of calculation abilities and visualize it into three dimensions for concrete measurement, thus performing a standardized proficiency test of whether students’ calculation abilities have changed. As previously highlighted by Commodari and Di Blasi (2014), any computation requires a certain level of calculation methods and knowledge and emphasizes the ability to perform calculations quickly and accurately.

In this study, three sub-abilities of calculation abilities as Figure 1 shows (calculation speed, calculation accuracy, and numerical knowledge) were used as the theoretical framework for a more accurate measurement of students’ calculation abilities. AC MT 6-11 thus was adopted as an instrument of this study.

![Figure 1. Framework for Calculation Abilities](image)

3. Methodology

3.1 Procedures and ethical measures

The study was conducted in a school in Shenzhen, a southeast city in Mainland China covering Grade 1 to 9, and consent were obtained from the participating students and their parents, the mathematics teacher, and the school principal before the study was...
implemented. Since there are different classes for each grade in this school, the study used cluster sampling, with one class in grade 4 randomly selected as an experimental group (n=40). For the control group, another class taught by the same mathematics teacher in the same grade was selected as the control group (n=38). A total of 78 students participated in the study.

In this study, a pre-and post-test quasi-experimental approach was used, as shown in Figure 2. Three days before starting and after ending the intervention, all students took a pre-test and post-test of the calculation test (i.e. AC MT 6-11). A 40-minute teacher training was conducted before the pre-test to ensure that the teacher understood the purpose and principles of the study. Both the pre-test and post-test were completed in classrooms at the targeted school, with proctoring and marking completed by a mathematics teacher in the same grade level who was not teaching in the participating classes, and a 20-minute teacher training was held with the marking teacher to agree on the scoring requirements and rules for the calculation test before scoring the test.

Before the main study, two additional teachers and six students from grade 4 were invited to participate in the pilot study to collect their opinions on the research instrument.

3.2 Intervention

3.2.1 The adopted mathematical game application

“Oral Math Hero” was adopted in the intervention of this study, and it was designed for elementary school students to practice mathematics in a fun way. As shown in Figure 3, students could choose the range of calculations and operations they want to practice. The game was based on players playing as different heroes, who need to fight monsters by answering questions then they got progressively harder. Students who gave correct answers could earn gold coins, and five consecutive correct answers unleashed a big move to defeat the monster. Gold coins could be used to buy higher-level hero characters. After each “battle” (stage), there was task feedback and a leaderboard to motivate students.
3.2.2 Pedagogy

The experimental group experienced the intervention which was based on the pedagogical model of Play Curricular Activity Reflection Discussion (PCaRD) (Foster & Shah, 2015). The PCaRD model consists of four phases (gameplaying phase, classroom activity phase, reflection phase, and discussion phase) to help teachers effectively integrate digital games into their instruction in the classroom (Denham, 2019). During the month-long game-based intervention, students in the experimental group were required to engage in a 15-minute activity of the mathematical game “Oral Math Hero” three times a week on iPads provided by the school, while the control group maintained the traditional classroom practice activities without learning with the mathematical game application.

Before the intervention, the researchers provided clear instructions to the teachers, who designed lesson plans based on the PCaRD model, including the objectives, content, and game-based activities. During the gameplaying time, every student learned with an iPad, and during each 40-minute session, the teacher briefly delivered 20 minutes of mathematics lesson content and divided the subsequent 15 minutes into four sections based on the PCaRD model, as shown in Figure 4.

Figure 4. Intervention based on PCaRD model.

3.2.3 Calculation practice in the control group

To gain a more comprehensive understanding of the impact of the intervention on calculation abilities, while the experimental group implemented the classroom intervention of the
mathematical game “Oral Math Hero”, the control group implemented the traditional calculation practice method, which is based on paper and pencil training, and performed oral calculation practice or calculation training in exercise books at the same frequency and duration as the experimental group.

3.3 Instrument

As stated, Abilita diCalcoloz Calculation Abilities Memory and Training Group 6-11 (AC MT 6-11) was used as a calculation abilities test in this study, which was a standardized calculation test that requires students to determine if students’ calculation abilities levels changed before and after the four-week intervention and were expected to take 40 minutes. The calculation abilities test included scores in three dimensions: numerical knowledge, calculation speed, and calculation accuracy (Cornoldi et al., 2002). There were eight tasks in this calculation abilities test: Written Calculation; Size Comparison, Word–Number Transcoding, Numerical Facts, and Number Ordering; Mental Calculation; Written Calculation2; Enumeration; and Numerical Facts (Cornoldi & Lucangeli, 2004). According to Cornoldi and Lucangeli (2004), the calculation abilities test is a psychologically trustworthy assessment of calculation abilities, and the sources of scores for the three dimensions of calculation abilities were shown in Table 1.

Table 1. Sources of Scores for Different Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation Speed</td>
<td>Time for answering Written Calculation, Mental Calculation, Enumeration, Written Calculation2</td>
</tr>
<tr>
<td>Numerical Knowledge</td>
<td>Combined score: Size Comparison, Word–number transcoding, Numerical facts, Number ordering</td>
</tr>
<tr>
<td>Calculation Accuracy</td>
<td>Total Error Scores</td>
</tr>
</tbody>
</table>

3.4 Data analysis

The pretest scores of the experimental and control classes in both grades were subjected to separate independent sample T-tests to ensure that there was no significant difference in the basis of calculation abilities between the two classes. After that, a single-factor analysis of variance was conducted for all pretest and post-test scores. Therefore, ANOVA (including Levene’s test of the quality of error variances) with a confidence interval of 0.95 is chosen to eliminate the effect of the intervention on the results and was followed by a normal distribution and homogeneity of the regression slopes to ensure that there was no interaction between the pretest scores and the intervention (“Oral Math Hero” game application). The above data are analyzed using IBM SPSS 28.0.

4. Results and Findings

4.1 Descriptive analysis

Cronbach's alpha was calculated using the pre-test results to ensure the reliability and reliability of the test before conducting the post-test. The reliability analysis showed that Cronbach's alpha was greater than 0.7 for all factors ($\alpha_{CS} = 0.727$, $\alpha_{CA} = 0.841$, $\alpha_{NK} = 0.769$), confirming that the standard calculation abilities test is a reliable research instrument. The results of the independent samples T-test ($p=0.930<0.05$) showed that there was no significant difference in the pre-test scores of the experimental and control groups in grade
which meant that the two groups had a similar basic level of calculation abilities before the start of the study. The normality test was performed on the pre-test and post-test data of the calculation abilities test, and since the sample size of each group was <50, the results according to the Shapiro-Wilk test (p_{Pre-CS} = 0.056; p_{Post-CS} = 0.095; p_{Pre-CA} = 0.154; p_{Post-CA} = 0.085; p_{Pre-NK} = 0.280; p_{Post-NK} = 0.050) were all greater than 0.05, so the collected data conformed to a normal distribution. The homogeneity of variance was tested using Levene’s test, and the results (p_{Pre-CS} = 0.427; p_{Post-CS} = 0.882; p_{Pre-CA} = 0.237; p_{Post-CA} = 0.740; p_{Pre-NK} = 0.573; p_{Post-NK} = 0.622) suggest that the differences in the overall variance of the groups are small.

The descriptive analysis of the pre-test and post-test results of the calculation abilities test was shown in Table 2. Students in the experimental group demonstrated a significant increase in calculation accuracy and calculation speed, whereas students in the control group demonstrated a slight increase in calculation accuracy but a decrease in calculation speed and numerical knowledge compared to the pretest. Meanwhile, students in the experimental group drastically dropped their numerical knowledge.

### Table 2. Descriptive Statistics of Calculation Abilities Test in Pretest and Posttest

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Pretest</th>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Deviation</td>
<td>Mean</td>
<td>Std. Deviation</td>
</tr>
<tr>
<td><strong>Calculation Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>3.88</td>
<td>3.560</td>
<td>6.03</td>
<td>11.173</td>
</tr>
<tr>
<td>Control Group</td>
<td>6.42</td>
<td>9.081</td>
<td>6.58</td>
<td>11.387</td>
</tr>
<tr>
<td><strong>Calculation Speed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>15.48</td>
<td>2.828</td>
<td>16.18</td>
<td>2.688</td>
</tr>
<tr>
<td>Control Group</td>
<td>18.34</td>
<td>1.805</td>
<td>17.00</td>
<td>2.857</td>
</tr>
<tr>
<td><strong>Numerical Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>38.58</td>
<td>5.679</td>
<td>33.55</td>
<td>0.959</td>
</tr>
<tr>
<td>Control Group</td>
<td>39.18</td>
<td>3.303</td>
<td>32.55</td>
<td>2.177</td>
</tr>
</tbody>
</table>

4.2 Analysis of variance (ANOVA)

According to the results of the T-test, there was no significant difference between the pretest scores of the experimental group and control group (p=0.930>0.05), indicating that the two groups of the fourth grade had similar bases. Therefore, a single-factor ANOVA was conducted to investigate whether there was a significant difference between the data of the experimental and control group.

### Table 3. ANOVA results of Calculation abilities in Experimental and Control Group

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>p</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall Calculation Abilities</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>2.708</td>
<td>0.004*</td>
<td>126.323</td>
</tr>
<tr>
<td><strong>Calculation Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>3.908</td>
<td>0.004*</td>
<td>137.894</td>
</tr>
<tr>
<td><strong>Calculation Speed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>28.174</td>
<td>0.001*</td>
<td>160.109</td>
</tr>
<tr>
<td><strong>Numerical Knowledge</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>6.972</td>
<td>0.063*</td>
<td>19.385</td>
</tr>
</tbody>
</table>

The results (Table 3) showed that overall speaking, students in the experimental group had significantly higher calculation abilities than those in the control group (p=0.004).
Specifically, among the three dimensions of calculation abilities, calculation accuracy (p=0.004) and calculation speed (p=0.001) were significantly different between the two groups in the posttest. Therefore, it was possible to draw a conclusion that digital game-based learning with Oral Math Hero on iPad significantly improved the calculation abilities of the experimental group of grade 4 students, especially the students’ calculation speed and calculation accuracy. However, no significant difference was found between the students’ ability in the experimental and control classes in terms of numerical knowledge (p=0.063). Learning with this mathematical game application might not lead to acquiring more mathematics knowledge.

5. Discussion and Conclusion

Calculation accuracy, calculation speed, and numerical knowledge as three objective measurements were considered to measure the impact of digital game-based learning with a mathematical game application on the calculation abilities of grade 4 students (Cornoldi & Lucangeli, 2004). Specifically, we predicted that DGBL with the mathematical game application, in terms of calculation accuracy, calculation speed, and numerical knowledge, would be as effective as normal mathematics class activities in improving the calculation abilities of grade 4 students, and the experimental group would perform better than the control group. Consistent with this hypothesis, the results of the post-test showed that the experimental group had significantly higher overall calculation abilities test scores and scores in terms of calculation accuracy and calculation speed than the control group. Such findings generally confirmed some studies on the use of mathematical games in mathematics classrooms (e.g., Núñez Castellar et al. 2014; van der Ven et al. 2017; Castellar et al. 2015), suggesting that DGBL with the mathematical game can improve certain mathematical skills by increasing students’ calculation accuracy and speed of mental computation. The findings are also in line with those of Miller and Robertson (2011) that elementary school students who played mathematical games had twice the calculation speed of control group students who did not play.

However, an unexpected finding occurred in this study: the numerical knowledge score gained in the post-test was lower for the experimental group than in the pre-test. More specifically, the findings of this study did contradict some previous studies (Brezovszky et al., 2019) that DGBL with mathematical game application did not improve students’ numerical knowledge significantly. Three potential explanations for the observed phenomenon, each of which is not exclusive of the others. Firstly, it is plausible that the “Oral Math Hero” game did not affect numerical knowledge to a significant extent since it provided tasks of a less complex and purely arithmetic nature. Secondly, students might not think deeply about certain calculation rules, such as the multiplication exchange law or the distributive law, in order to complete the task quickly in the exercise. Thirdly, the superior calculation methods were not presented in the task feedback or the answer situation screen of the game, so students had no chance to achieve more numerical knowledge in this game.

There were several limitations to this study. Firstly, although we have explored the effects of DGBL on calculation abilities by standardized calculation abilities tests, how calculation abilities are altered by digital games, how it varies with the type and content of digital games, and how the afterschool environment may affect calculation abilities remains unanswered. Future research would be expected to examine how DGBL affects students’ calculation abilities, taking the afterschool environment and the type and content of games into consideration. Secondly, the duration of the intervention (four weeks, 2-3 times per week) was relatively short and it was only used in the classroom, so its impact on students’ calculation abilities might have been limited. In addition, the duration of this study lasted only one month, which enabled the study to explore the short-term impact of DGBL with mathematics games on students’ calculation abilities but lacked a longer period of investigation to explore whether this impact would change according to the time change. Longitudinal trials and tracking of students’ calculation performance over time might be used to offer an understanding of potential long-term impacts on students’ calculation abilities.
Thirdly, the sample size of this study was relatively small, only 78 students, and the fact that all data were collected through calculation abilities tests in a mainstream school in Mainland China might have led to an under-representation of the study, limiting the generalizability of the findings to other educational settings or groups of students. In addition, the participants of this study were mainly focused on grade 4 students, with no attempt to cover the entire elementary school level or include students’ and teachers’ interviews, lacking multiple data sources to provide triangulation. A study on a larger scale involving after-school interventions in more districts with more sources of data such as classroom observations and individual interviews could be considered. Fourth, the study concentrated on the influence of DGBL with a mathematical game on students' calculation abilities while ignoring the impact of the game on students' basic mathematical skills, such as problem-solving, spatial thinking, numeracy, and sense of amount, as well as their motivation. Future research can investigate whether mathematical game impact students’ motivation and other components of mathematics learning.

In conclusion, this study showed that DGBL with a mathematical game application, compared to traditional calculation practice, could improve the calculation abilities of fourth graders, especially in terms of calculation speed and accuracy. Nevertheless, to present a comprehensive picture of how learning with a mathematical digital game affect calculation abilities, the data of the second-stage qualitative research is now analyzed to explain in what ways students were affected by game settings or elements. After that, implications for future DGBL with a mathematical game application can be proposed and all these will be presented in a forthcoming paper.

Acknowledgments

We would like to thank all the people who prepared and revised previous versions of this document.

References


The effects of video game playing on attention, memory, and executive control—ScienceDirect. (n.d.). Retrieved 5 February 2023


Develop and validate STEM education activities using the “6E Design Teaching Model”: Taking “Dynamics and Energy Conversion in Sail Car Design” as an example

Lu CHEN\textsuperscript{a}, Yang YANG\textsuperscript{b} & Guang CHEN\textsuperscript{c}

\textsuperscript{a}Faculty of Psychology, Beijing Normal University, China
\textsuperscript{b}School of Educational Technology, Faculty of Education, Beijing Normal University, China
\textsuperscript{c}School of Educational Technology, Faculty of Education, Beijing Normal University, China

*guang@bnu.edu.cn

Abstract: This study employed a mixed research approach to analyze commonly used teaching models in STEM education and integrated the findings of comprehensive STEM research. The 6E Design Teaching Model (Engage, Explore, Explain, Engineer, Enrich and Evaluate) was selected, following the “Science-Technology-Engineering-Mathematics integration” approach. It was used to develop STEM education activities targeting students in the 5th and 6th grades, focusing on “Dynamics and Energy Conversion in Sail Car Design.” Through three rounds, the effectiveness and feasibility of the activities were examined. The study employed quantitative tools and qualitative interviews (including teachers and students) to improve the teaching process and instructional framework. The results indicate that (1) STEM education activities developed using the 6E Model significantly enhance students’ interest in engineering subjects when applied in teaching practice; (2) the instructional activities can effectively be applied in classrooms that integrate scientific inquiry and engineering practices. Over the course of the three rounds, the participating teachers demonstrated improved classroom control and time management, while student scores in engineering design projects showed noticeable progress. Through the optimization of the teaching process, instructional framework, and other aspects such as materials, operations and management, the study ultimately developed a highly practical STEM education case that integrates engineering practices into science inquiry classrooms. This case serves as a reference for frontline science teachers and researchers.

Keywords: 6E Design Teaching Model, Integrated STEM Education, Activity Design

1. Introduction

STEM education (Science, Technology, Engineering and Mathematics) is an important means to cultivate students’ innovation and practical abilities. Although researchers have different understandings and perspectives on the concept, connotation and characteristics of STEM education(Hubelbank & Oliva, 2014; Mersin National Education Directorate, Turkey & Uyguş, 2022), they all point to the common goal of interdisciplinary integration, with a broader and wider scope of integration(Radloff & Guzey, 2016). Various studies have shown that among the four disciplines of S, T, E and M, mathematics and science are the primary subjects, while others serve as connecting links, resulting in the marginalization of the engineering discipline. In recent years, many countries have implemented engineering education reforms in K-12 (kindergarten through twelfth grade) schools(Moore et al., 2014; National Research Council, 2009). For example, a report titled “Preparing and Inspiring the Future STEM Workforce of America: K-12 Science, Technology, Engineering and Mathematics (STEM)” submitted to the President of the United States emphasized the need
to cultivate students with a strong STEM background to remain competitive in the global society (National Research Council, 2011). Integrating scientific inquiry and engineering practices into science classrooms to form Integrated/Integrative STEM Education has become one of the important trends in the development of STEM education worldwide (Guzey et al., 2016). Therefore, our research will follow six steps (see Figure 1).

![Figure 1. 6E Design Teaching Model (the 6E Model)](image)

Researchers divided the overall framework of integrated STEM education into four components: goals, outcomes, nature and scope of integration and implementation (National Research Council, 2014). This study adopts an integrated approach based on scientific exploration and engineering design, to stimulate students' interest in exploration, encourage them to design and create engineering projects and apply interdisciplinary knowledge.

The 5E teaching model was developed by the Biological Sciences Curriculum Study (BSCS) in the 1980s in the United States based on the constructivist learning perspective for scientific inquiry. It includes five stages: engagement, exploration, explanation, elaboration, and evaluation to form an understanding of scientific concepts. It has been widely used in biology, physics, chemistry, science and out-of-school science education activities (Duran & Duran, 2004). Barry (2014) proposed adding an “engineering” stage to the 5E teaching model, resulting in the 6E Model: Engage, Explore, Explain, Engineer, Enrich and Evaluate. Engage aims to stimulate interest and enhance learning participation, explore provides opportunities for constructing self-knowledge, understanding analysis and hands-on operation, explain provides opportunities for reflection, explanation, modification and elaboration of knowledge, engineer provides opportunities to apply learned knowledge and technology to practice, enrich promotes in-depth exploration of knowledge and evaluate checks and evaluates learning outcomes.

Khaeroningtyas et al. (2016) developed a STEM education activity with the theme of “Temperature and Its Changes” using the 6E Model. They implemented a quasi-experimental design and found that the experimental group achieved higher scores in the topic of temperature and its changes. Lai & Chu (2017) utilized the 6E Model to develop a STEM education activity with the theme of “Quadcopter Assembly”. The learning performance of 48 middle school students showed significant improvement. Hsiao et al. (2022) implemented a quasi-experimental design found that using robot-based practices to develop an activity that incorporated the 6E model can improve elementary school students’ learning motivation, learning performance, computational thinking ability, and hands-on ability. The STEAM education activities, using the 6E Model, were implemented with third-grade students in Thailand. After 12 class hours, the students found improvements in academic performance, scientific creativity, and innovation abilities (Jongluecha & Worapun, 2022). And the 6E Model in a practical activity involving the design of egg protection devices was found to have a positive impact on the technological attitude and technological inquiry abilities of the participating secondary school students (Lin et al., 2020).
Two study hypotheses were proposed: utilizing the 6E Model to develop STEM education activities can be effectively applied in teaching practice and STEM education activities developed using the 6E Model can enhance students' interest in engineering learning. This study employed the 6E Model to develop and design integrated STEM education activities and validated their effectiveness through implementation, resulting in a curriculum design method of practical significance.

2. Method

Data was collected using quantitative tools and qualitative interview methods. Quantitative data included scores of students' engineering design projects and surveys on their interest in engineering and science subjects. Qualitative data included interviews with teachers and students. The research process is as follows (see Figure 2).

2.1 Participant

This study recruited participants through collaboration with one university community, a tutoring organization and a parents' group. A total of 56 students (15 in 5th grade, 40 in 6th grade) and one teacher who was a science education instructor from the tutoring organization.

![Figure 1. Research Process Flowchart](image-url)
2.2 Research Tools

2.2.1 The STEM Interest Survey

The STEM Interest Survey is an 25-item self-report scale with 5 factors of interesting in science, technology, mathematics, engineering and STEM careers (Tyler-Wood et al., 2010). Items were answered on a 7-point Likert scale that ranges from 1 (not at all important) to 7 (extremely important). The Cronbach's Alpha is between 0.84 and 0.93.

2.2.2 Student Work Scoring Sheet

The the Performance-Based Evaluation(PBE) are four factors of completion, feasibility, aesthetics and innovation (Jin et al., 2015). The PBE uses a 4-point scale (1-4), with high stability and reliability.

2.2.3 Interview Outlines

The study includes interviews with teachers and students. Teacher interviews primarily focus on the situations and issues encountered during actual teaching processes, covering topics such as teaching methods and evaluation methods. For example, do you consider this teaching method to be similar or different from previous methods? What are the advantages and disadvantages? Student interviews aim to explore the application of scientific principles and mathematical knowledge in engineering practices during the activity, as well as the difficulties encountered in the classroom and overall satisfaction with the course activities. For example, what difficulties did you encounter during the learning process? For example, abstract concepts or challenging hands-on operations. Please provide specific examples.

3. Research Steps

3.1 Development of Teaching Activities

In the development of the 6E Model, five aspects were considered: needs analysis, content design, teaching resources, process design and evaluation of outcomes. The needs analysis involved analyzing educational resources and learners. The study adopted classroom-based activities with a relatively short duration and low resource requirements. Fifth and sixth-grade students were chosen over third and fourth-grade students due to their higher levels of self-management skills, hands-on practical abilities and proficiency in mathematical operations and independent thinking.

Content design involved conducting surveys among frontline teachers and curriculum design engineers, followed by preliminary analysis of potential themes. Ultimately, ancient Chinese transportation vehicles, specifically the sail-powered cart, were selected as the theme for the curriculum design.

Teaching resources included the development of teaching handbooks, student handbooks and activity materials kits. The teaching handbook guided teachers in understanding the curriculum design concepts, teaching processes and lesson preparation. The student handbook helped students plan their learning activities, record important course information, and reflect on their experiments. The materials kit bundled together all the materials used in the curriculum and was distributed to the students. The process design stage consisted of the following phases: engage, explore, explain, engineer, enrich and evaluate. Outcome assessments included measuring in student interest in engineering learning, grading engineering design projects and interviewing teacher and students.

3.2 Implementation of teaching activities
A total of three rounds of courses were conducted, with one round per week, lasting 120-180 minutes each time. After each round of courses, the researchers conducted statistical analysis on the data and conducted interviews with the teachers and students for further improvement. The detailed process of the first round of courses is as follows (see Table 1).

<table>
<thead>
<tr>
<th>Phase</th>
<th>Time(mins)</th>
<th>Detailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engage</td>
<td>20-30</td>
<td>Teacher’s explanation and student discussions enable students to learn about the history of sailboats and the science and applications of sail design.</td>
</tr>
<tr>
<td>Explore</td>
<td>20-30</td>
<td>Students participate in a sailboat-making competition, followed by sharing their creations, after which the teacher provides explanations of the underlying principles, such as friction and material selection.</td>
</tr>
<tr>
<td>Engineering</td>
<td>20-40</td>
<td>Students redesign their sailboats, with the teacher introducing the concepts of engineering and guiding them through the engineering design process to optimize their designs.</td>
</tr>
<tr>
<td>Enrich</td>
<td>40-50</td>
<td>Students work in groups of four to collaboratively design and build sailboats for a competition.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>20-30</td>
<td>The teacher provides feedback and students share their experiences, with the teacher concluding the activity with a summary.</td>
</tr>
</tbody>
</table>

3.2.1 Improvement of Teaching Process

In the second round of activities, the engage was conducted by group teaching. The topic focused on common tools that utilize wind energy and interactive discussions were held with the students to introduce the concept of harnessing wind energy. In the exploration phase, the straws were replaced with a more rigid and easily fixable material and the structure-building component was simplified to using foam boards for support. The explanation replaced the traditional lecture-style delivery of key points with interactive mini-games. The teacher used the physical model of the cart to ask questions and engage the students in discussions while recording their responses. In the engineering, more time was allocated for teacher explanations and student hands-on activities. During the enrich, students formed teams of four to participate in a competition. In the evaluation, each student conducted a brief analysis.

After optimizing these changes, the subsequent recruitment and classes were conducted, and the teachers were able to smoothly control the classroom. The teaching experience improved and there was an increase in the completion rate of student projects.

3.2.2 Improvement of Teaching Framework

After the first round of activities, exploration and engineering components have been added to provide reference experimental procedures for students who may encounter difficulties in certain operations (More see appendix1, appendix2). With the inclusion of experimental procedures, in the second round of activities, over 50% of students successfully completed 60% or more of the content and design of their projects, as outlined in the learning manual. For the third round of activities, the SCAMPER(Substitute, Combine, Adapt, Modify/Magnify, Purpose, Eliminate and Rearrange/Reverse) brainstorming cognitive tool was used to guide students through the problem-solving process(Serrat & Serrat, 2017). Additionally, the researchers employed the 3-2-1 Quick Assessment Method (3-2-1) commonly used in the 5E inquiry-based teaching model, to encourage students to share three takeaways, two challenges and one aspect they found particularly effective during the entire activity.
Following these improvements, the success rate of student exploration experiments and the completion level of their projects significantly increased in the third round of instruction.

3.2.3 Other Improvements

Firstly, in terms of personnel coordination, there was enhanced communication among researchers, recruitment personnel and teachers. The recruitment process was coordinated and planned to avoid blind enrollment. Secondly, in terms of course materials, canvas, thinner wood, and additional small wooden sticks were added to meet the students' needs and enthusiasm for fabrication and design.

With these improvements, the recruitment of students became more targeted, avoiding the interference of ineffective samples. The use of additional materials by students led to more refined and complex design projects, resulting in improved performance during testing.

4. Analysis of STEM Learning Interest

4.1 STEM Learning Interest Analysis Results

Through three rounds of recruitment and course implementation, a total of 27 valid STEM interest survey forms were collected (6 forms in the first round, 10 forms in the second round and 11 forms in the third round). The analysis results are as follows (see Table 2).

<table>
<thead>
<tr>
<th></th>
<th>Science</th>
<th>Math</th>
<th>Technology</th>
<th>Engineering</th>
<th>STEM Career Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First round</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.00</td>
<td>1.50</td>
<td>7.67</td>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td>df</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
</tr>
<tr>
<td>t</td>
<td>1.94</td>
<td>1.31</td>
<td>4.97</td>
<td>0.54</td>
<td>1.23</td>
</tr>
<tr>
<td>P</td>
<td>0.11</td>
<td>0.25</td>
<td>0.004*</td>
<td>0.61</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>Second round</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>0.27</td>
<td>1.36</td>
<td>3.09</td>
<td>0.00</td>
<td>0.27</td>
</tr>
<tr>
<td>df</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>t</td>
<td>0.40</td>
<td>1.04</td>
<td>3.40</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>P</td>
<td>0.70</td>
<td>0.33</td>
<td>0.007*</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Third round</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.50</td>
<td>0.70</td>
<td>5.60</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>df</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
<td>10.00</td>
</tr>
<tr>
<td>t</td>
<td>2.24</td>
<td>0.90</td>
<td>3.89</td>
<td>0.92</td>
<td>2.25</td>
</tr>
<tr>
<td>P</td>
<td>0.05</td>
<td>0.39</td>
<td>0.004*</td>
<td>0.38</td>
<td>0.05</td>
</tr>
</tbody>
</table>

It can be observed that there were no significant differences in students' interest in science, technology, mathematics, and STEM careers. However, there was a significant difference in engineering learning interest after instructional intervention.

4.2 Analysis Results of Engineering Design Project

The research conducted a scoring and recording of the engineering design projects in each round. There were 24 projects in the first round, 19 projects in the second round and 12 projects in the third round. In each round, the final scores for each dimension of the projects
were calculated as the average scores. The analysis focused on comparing whether there was an improvement in the scores of student projects in terms of completeness, feasibility, aesthetics and innovation (see Figure 3.).

![Graph showing analysis results]

*Figure 2. Analysis Results of Engineering Design Project*

During the three rounds, there was an improvement in the completeness, feasibility and innovation of student engineering design projects. The aesthetics also showed improvement in the third round, although the difference was not significant. Moreover, there was a noticeable improvement in the completeness between the first and second rounds and the innovation significantly improved after the second round (More see appendix3, appendix4, appendix5).

### 4.3 Interview analysis results

From an overall perspective, after the first round of the activities, students who were interviewed (both those who successfully completed the engineering design projects and those who did not) could only recall the data mentioned by the teacher in class and provide a simple explanation without being able to articulate how to apply it to the design of the sail car. However, in the second and third rounds, some students were able to assess and evaluate relevant engineering knowledge and concepts. In the third round, some students even proposed incorporating knowledge of buoyancy to make the sail car amphibious. Furthermore, in terms of satisfaction with the course and interest in the activities, apart from some students in the first round expressing a shortage of materials, overall satisfaction was high.

### 5. Discussion

The suggestions and improvement ideas you have provided are intended for frontline teachers and external science educators to consider. Firstly, you recommend making thorough preparations when implementing activities using the 6E Model. In the engage, you can organize the curriculum around real-life engineering problems that need solving and design inquiry experiments with moderate difficulty to provide necessary scaffolding for students, thereby stimulating their confidence in hands-on design. In the explanation, teachers can conduct inquiry experiments,
explain in detail the scientific concepts, principles and core knowledge required to solve engineering problems by referring to relevant materials (including interdisciplinary subjects such as science history, technological development, social sciences, mathematics, etc.). During the engineering, teachers can explain the concepts and design processes involved in engineering and utilize creative thinking tools like SCAMPER brainstorming to assist students in organizing and conceptualizing their design ideas. Furthermore, teachers should personally test and create the products that students are expected to make in class, engage in iterative testing and provide timely feedback and suggestions. In the evaluation phase, employ diverse assessment methods, avoid solely teacher-led evaluation and consider using tools such as the 3-2-1 to allow students to share their three takeaways, two questions and one positive aspect, enabling them to reflect on their learning outcomes and summarize the course content. Secondly, you emphasize the importance of teamwork and collaboration in the implementation process to enhance work efficiency and prevent errors. For example, in the case of the small sail car activity, the teaching instructor's expertise in physics and engineering knowledge, along with their extensive teaching experience, greatly aided the researcher in improving the instructional design.

Additionally, you highlight the challenges encountered during the implementation, including difficulties in student recruitment and limitations concerning experimental materials, tools, and facilities. Teachers may allocate a significant portion of their teaching time to maintain order during the process of students' hands-on construction, which results in less time dedicated to explaining the instructional content. You suggest that this issue may arise from insufficient consideration of the actual teaching situation and potential problems during the initial research phase. Therefore, when designing activities with engineering themes, you recommend that teachers pay attention to preparatory work, such as collecting materials, conducting model testing, evaluating effects, and providing abundant instructional scaffolding to maintain students' confidence in autonomous design.

6. Conclusion

Based on the research findings and experiences integrating comprehensive STEM education with K-12 engineering education, an exploration of incorporating engineering elements into science classrooms was conducted. A STEM education case study was designed and developed with a focus on the design of an ancient Chinese transportation vehicle called the sail car. It was discovered that when STEM education activities developed using the 6E Model were applied in teaching practices, they significantly enhanced students' interest in the field of engineering. The instructional activities proved effective for integrating scientific inquiry with engineering practices in the classroom. Through optimizing the teaching process, instructional scaffolding and other aspects such as materials and operations management, a highly practical STEM education case that integrates engineering practices into science inquiry classrooms was developed, offering a new approach for international STEM education curriculum development.

7. Prospects

The limitations of the study include focusing only on mechanics and “technology and engineering” themes without forming a complete series of courses, a small sample size and a single evaluation method. In future research, the enrichment could be expanded by incorporating the use of digital tools to guide students in adding sensors and data analysis tools that provide real-time monitoring and display of the car's motion status to the existing sail car model. In terms of curriculum topics, organizing instructional content around technology and engineering or broader themes could be considered to develop a series of courses.
References


Appendix

Appendix 1. Explore (added)

1. Cut the sail and base with precision.

2. Use a carving knife to create two holes of straw size on the top and bottom of the sail, and connect the straw with the sail.

3. Secure the straw on the cut-out foam board base with double-sided tape and nylon rope.

4. Reinforce the base and straw with double-sided tape. Voila! You now have a model that runs with just one breath.
Appendix 2. Engineering (add)

1. Strengthen and refine the design of the sail, connecting it to the mast and a base with small holes.

2. Measure a rectangle on a wooden board with dimensions of 20cm length and 60cm width for the sail car base. 

Cut it open with a saw following the guidelines. Sawing can be done by placing the board on the edge of a table, exposing a small portion of it, using one hand to press one side onto the table, and tilting the saw with the other hand to saw it back and forth.

3. Use a hot glue gun to attach three popsicle sticks into one long stick, then divide it into two segments of 3.5cm each and connect three sticks together.
4. Draw an isosceles trapezoid on both sides of the wood stick and cut off the excess with scissors. Repeat the same process to make another isosceles trapezoid of the same size and shape.

5. Use a hot glue gun to attach the sawed base to the trapezoid side of the sail car.

6. Use a hot glue gun to connect the triangle iron to the bottom of the car, and assemble the axle and wheels.
7. Secure the base of the sail on the front of the sail car using a hot glue gun.

8. The sail car is now complete. Design a logo and embellish it beautifully in preparation for a mineral water racing competition.

Appendix 3. The first round of student works (partial)
Appendix 4. The second round of student works (partial)

Appendix 5. The third round of student works (partial)
The Impact of Gamified Assessment on the Learning Burnout of Undergraduate Computing Students: a Quasi-experimental Research

Beilei ZHANG\textsuperscript{a}*, Alex Wing Cheung TSE\textsuperscript{b}
\textit{The University of Hong Kong, Hong Kong SAR}
\textsuperscript{a}u3598214@connect.hku.hk; \textsuperscript{b}awc@hku.hk

\textbf{Abstract:} Learning burnout refers to the negative attitude and behavior of being bored with learning due to learning pressure or lack of learning interests. It affects not only students’ academic performance but also their physical and mental health. Currently, the research on learning design to alleviate learning burnout has only been the subject of a few studies. Considering that gamified assessment theoretically gave it potential to affect students’ learning burnout, 120 undergraduate computing students at a university in mainland China participated in this study to investigate whether such a learning process was conducive to affecting learning burnout. The research was a quasi-experimental design with pretest and posttest. For the experimental group (n=60), gamification elements were integrated into two sessions including the question and answer sessions and after-class quizzes by Quizizz for six weeks (2 hours per week). During the data collection, all participants involved completed the pretest and posttest of the Maslach Learning Burnout Inventory-Student Survey via an electronic questionnaire. The quantitative data were analyzed through ANCOVA. This study found that students’ learning with gamified assessment for six weeks did contribute to alleviating students’ learning burnout, particularly in terms of cynicism and low efficacy. The findings can be a reference for exploring the integration of gamified assessment into tertiary education.

\textbf{Keywords:} Gamified assessment, learning burnout, gamification, quasi-experimental design

1. Introduction

With the widespread use of information technology in tertiary education, mental health risks, especially those related to learning burnout among undergraduate students, are particularly obvious (Zhao et al., 2022). Affected by learning burnout, students’ characteristics of poor learning performance become more common, including inattentiveness, low learning efficiency, high learning anxiety, and even resistance to learning. Some schools have put forward corresponding psychological counselling suggestions aiming at effectively alleviating the emotional problems of learning (Ates, 2016). Gamification can be defined as the use of game elements in non-game environments (Deterding et al., 2011). When gamification is applied to the educational area, it motivates and educates learners through appropriate game elements. Gamification has been exhibited as an effective approach for teaching computer courses in tertiary education (Ahmad et al., 2020). The use of gamification in educational contexts has led to the creation of new assessment methods. More specifically, game elements such as badges and levels are added to student
assessments. Past research has shown that gamification has been effectively implemented in a number of learning environments (Chung & Lin, 2022; Saleem et al., 2022). However, the use of gamification to assess student learning has only been the subject of a few studies (Zhang & Fang, 2019). Notably, little research has been done to investigate the effects of gamified assessment in the context of tertiary education.

This study aims to explore whether gamified assessment contributes to alleviating undergraduate computing students’ learning burnout, hoping to provide a reference for exploring the improvement of such a problem. A more thorough analysis of how gamified assessment affects multiple dimensions of learning burnout has been conducting in the second stage of the research project based on the qualitative findings which will be reported in another paper.

The research question of the present study: Does gamified assessment affect the learning burnout of undergraduate computing students?

2. Literature Review

2.1 Gamification in Education

Gamification is the use of game elements or game design techniques in non-game contexts, with the aim of introducing game elements without detracting them from realistic contexts (Werbach et al., 2012). Gamification has been shown to enhance learning (Andriamirisoa, 2018; O’Neill et al., 2018). It is an effective way of boosting learners to achieve learning outcomes, motivating learners, and facilitating learning (Göksun & Gürsoy, 2019). MacKinnon et al. (2015) conducted a quasi-experimental study to investigate the impact of gamification on students’ motivation, and their conclusions indicated that gamification significantly improved learning outcomes. The gamification elements can be progress bars, leaderboards and badges, employed as alternative or supplementary assessment methods to scores (Tsai et al., 2018). Numerous educational research has used gamification for the last decade to stimulate learning motivation, increase engagement and accelerate learning (Dicheva et al., 2015). Therefore, further exploration and research into the integration of gamification in education is necessary to fully realize its potential in enhancing the instructional effectiveness. In this study, gamification was integrated into the assessment.

2.2 Gamified Assessment

When the term “gamified” is applied to assessment, it describes a design process that incorporates gaming aspects to improve an existing assessment method (Georgiou & Nikolau, 2020). Gamified assessment can creatively apply gamification elements to related tasks. It gives the potential to promote the efficiency of students’ information exchange to achieve the purpose of promoting learning (Shute, 2008).

Various gamification elements can be incorporated into existing assessments, including progress bars, points, and badges (Landers et al., 2015). The implementation of gamified assessment is usually carried out by the teacher using some gamified assessment systems (Delacruz, 2011), such as Kahoot!, Socrative and Quizizz. Teachers can rapidly and effectively assess students’ learning with the aid of gamification systems and students can benefit from immediate feedback. A recent empirical study has further confirmed that technology-enhanced gamified assessment was more motivating than paper tests and improved academic performance (Zainuddin et al., 2020). Using gamification interaction could improve course completion, interaction and motivation rate (Hassan et al., 2021). Dias (2017) found that gamified assessment increased the proportion of students participating in the classroom. A literature review concluded that the majority of related papers have concluded that gamified assessment has positive effects on learning motivation, learning achievement, and learning satisfaction (Wang & Tahir, 2020).

In summary, gamified assessment is an innovative approach to assessment that incorporates game elements into educational contexts to positively motivate and engage
students in learning. Gamified assessment has the potential to affect the learning burnout of undergraduate students and enhance their learning effectiveness.

2.3 Learning Burnout

Learning burnout refers to the negative attitude and behavior of being bored with learning due to learning pressure or lack of learning interests (Meier & Schmeck, 1985). Burnout among students can have a negative impact on their academic achievement and personal well-being (Ribeiro et al., 2018). These findings suggest that addressing learning burnout is crucial for students’ overall health as well as for their academic achievement.

A study reported that learning motivation and burnout have a significant negative correlation (Cazan, 2015). Current relevant research focuses on exploring the influencing factors of learning burnout from a psychological perspective and seeking intervention measures. An empirical study found that teachers’ emotional support has a mitigating effect on learning burnout (Zhao et al., 2018). Based on existing relevant research, a lack of motivation is one of the causes of burnout (Felaza et al., 2020). Gamified assessment has been shown to stimulate motivation through the incorporation of gamification elements, immediate feedback, progress tracking, and healthy competition among students (Oliver, 2017; Pitoyo & Asib, 2020). Hence, the gamified assessment gives it potential to alleviate learning burnout considering learning motivation. Learning anxiety is an essential factor influencing learning burnout (Guangwen et al., 2013). The application of gamified assessment in traditional classroom teaching can effectively alleviate learners’ learning anxiety (Lee et al., 2019). Hence, gamified assessment may indirectly help to alleviate learning burnout. Previous research has demonstrated that gamification can give learners pleasant psychological experiences like immersion and flow experiences, which could promote learning (Pitoyo, 2019; Watson et al., 2011). Gamification also places learners in a positive psychological state of learning, which has a positive impact on learners’ academic achievement and engagement, thus alleviating learning burnout (Göksün & Gürsoy, 2019).

The above literature review highlights the negative relationship between learning motivation and burnout, giving gamified assessment the potential to affect learning burnout. The review also suggests that gamified assessment has the potential to alleviate learning burnout by stimulating motivation and promoting positive psychological experiences. This provides a basis for exploring if gamified assessment can affect the learning burnout of students. This study investigated such an issue among the computing undergraduates.

3. Methods

3.1 Theoretical Framework, Instrument and Intervention

Learning burnout can be measured with the Maslach Burnout Inventory-Student Survey, including three dimensions: emotional exhaustion, cynicism and low efficacy (Schaufeli et al., 2002). The scale is administered on a seven-point Likert scale ranging from 1 (absolutely inappropriate) to 7 (absolutely appropriate). Higher scores indicate higher levels of learning burnout. The study kept the structure of the original scale by referring to the Maslach Burnout Inventory-Student Survey. The questionnaire consists of 15 items divided into three dimensions above. Five items are designed to measure emotional exhaustion, which describes feelings of emotional depletion due to studying-related stress. Cynicism is measured with four items that assess negative attitudes and detachment towards studying, classmates, or teachers. Lastly, six items are employed to evaluate low efficacy, which refers to a diminished sense of competence and accomplishment in the realm of academics. The Learning Burnout Scale was translated into Chinese by a professional with expertise in educational psychology. Participants were provided with instructions in their native language and encouraged to seek clarification if needed. A pilot study was conducted to establish the validity and reliability of the translated version.
Empirical validation is necessary to confirm the hypothesized potential benefits of gamified assessments in reducing learning burnout. Nevertheless, the majority of research on learning burnout has been focused on medical students, which limits the potential for generalization of the observed findings (Almutairi et al., 2022; Fiorilli et al., 2022; Gradiski et al., 2022). All these points reveal the importance of this research. Based on the discussion above, the intervention in this study was guided by the theoretical framework presented in Figure 1 (Dicheva et al., 2015; Schaufeli et al., 2002).

![Figure 1. Alignment between the Theoretical Framework and Intervention](image-url)

The targeted participants were recruited from 120 sophomores who majored in computing at a university in mainland China. There are two groups, including the experimental (n=60) and control groups (n=60). A weekly two-hour Java programming course was provided with or without gamified assessment for six weeks. The same computing instructor taught two classes under a similar instructional schedule for the purpose of controlling the conditions. However, the Java programming curriculum for the two groups had different instructional designs. The instructor taught the control group based on the traditional paper-and-pen assessment. Whereas for the experimental group, students learned under gamified assessment during class and finished gamified after-class quizzes by Quizizz. Overall, the experimental group was delivered instructions with the gamified assessment, while the control group continued to use the traditional assessment. The assessment tasks were balanced in terms of content and level of difficulty. Though learning with different approaches, participants in both the control and experimental groups received the same type of feedback, which included correct answers and explanations for each item after completing the assessment. One of the researchers was in charge of the gamified learning environment support, gamified learning design, and teachers training.

The research activities took place over a seven-week period and included preparation, intervention and assessment activities. Weekly-2-hours-activities of classes with gamified assessment for six weeks were designed by referring to the theoretical framework (Figure 1). For the experimental group, the students were introduced to the concept of gamification, the relevant platforms and the game elements in the first week. From week two to week seven, an intervention of teaching and learning activities took place and instructions were given on different topics of Java programming. The entire class was conducted in a gamified environment. While the control group was taught traditionally without gamified assessment. At the end of each lesson, students would take an after-class quiz.

### 3.2 Research Design and Data Collection Procedures
A quantitative approach with a pretest-posttest quasi-experimental design was used in this study (Figure 2). The implementation of gamified assessment consisted of two components: gamified question and answer sessions during class and gamified quiz sessions after class. Quantitative data on burnout were collected by the Learning Burnout Scale (Schaufeli et al., 2002). Shortly before and after the intervention, the experimental and control groups took pretests and posttests of learning burnout. Averagely, it took about 5 minutes for participants to complete the questionnaires once, which were presented in electronic forms. During the data collection, students were given the e-questionnaire, and all the data was imported into SPSS as text files for data analysis.

Figure 2. The Overall Research Design

3.3 Gamified Assessment Instructional Design

3.3.1 Gamified Question and Answer Sessions

The design of gamified question and answer sessions is shown in Table 1 based on the design of gamified assessment by Werbach et al. (Werbach et al., 2012). Regarding the possibility of student choice, the principle of “freedom of choice” was applied in this study (Dicheva et al., 2015). The principle of “social engagement” including individual and team competitions (Li et al., 2014) was applied when designing gamified question and answer sessions. Students could find information on all these gamified components in Moodle, allowing them to track their performance and progress over time.

Table 1. The Design of Gamified Question and Answer Sessions (Werbach et al., 2012)

<table>
<thead>
<tr>
<th>Dynamics</th>
<th>Mechanics</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotions</td>
<td>Reward</td>
<td>Badges</td>
</tr>
<tr>
<td></td>
<td>Competition</td>
<td>Badges, Leader Board</td>
</tr>
<tr>
<td></td>
<td>Cooperation</td>
<td>Teams</td>
</tr>
<tr>
<td></td>
<td>Resource Acquisition</td>
<td>Points</td>
</tr>
<tr>
<td>Progression</td>
<td>Feedback</td>
<td>Points, Badges, Leader Board, Levels</td>
</tr>
<tr>
<td></td>
<td>Reward</td>
<td>Badges</td>
</tr>
<tr>
<td></td>
<td>Resource Acquisition</td>
<td>Points</td>
</tr>
<tr>
<td></td>
<td>Feedback</td>
<td>Points, Badges, Leader Board, Levels</td>
</tr>
<tr>
<td>Relationships</td>
<td>Cooperation</td>
<td>Teams</td>
</tr>
</tbody>
</table>

The five types of badges, “Active badge”, “Courage badge”, “Well done badge”, “Challenge badge”, and “Cooperation badge”, were awarded to students based on their performance in learning (see Table 2 for detailed design). Each badge represented a particular achievement; students could collect them as they progressed through the assessment. The gamified assessment system implemented a point-based reward structure, providing immediate feedback for correct answers and thus promoting accuracy. This incentivization, combined with the opportunity for students to form collaborative teams,
facilitated not only individual learning but also collective achievement through the accumulation of points and badges. The visible leaderboard fostered a competitive environment and further motivated students to excel. The multi-tiered structure of the assessment, which featured escalating levels of difficulty and reward, was aimed at engendering students a sense of accomplishment, thereby inspiring continued learning progression.

Table 2. Types and Descriptions of Badges Used in the Question and Answer Sessions

<table>
<thead>
<tr>
<th>Badge</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Active badge" /></td>
<td>Active badge</td>
<td>When a student takes the initiative to answer a question, he/she would be given an “Active badge”.</td>
</tr>
<tr>
<td><img src="image" alt="Courage badge" /></td>
<td>Courage badge</td>
<td>When a shy student answers a question for the first time, they would be given a “Courage badge”.</td>
</tr>
<tr>
<td><img src="image" alt="Well done badge" /></td>
<td>Well done badge</td>
<td>When a student gives a perfect answer, a “Well done badge” would be awarded.</td>
</tr>
<tr>
<td><img src="image" alt="Challenge badge" /></td>
<td>Challenge badge</td>
<td>Questions would be classified as easy, medium and difficult levels. The student would receive a “Challenge badge” when he/she completes a difficult question. One “Challenge badge” is equivalent to 2 “Active badges”.</td>
</tr>
<tr>
<td><img src="image" alt="Cooperation badge" /></td>
<td>Cooperation badge</td>
<td>During group tasks, if a group member shares a good suggestion for solving a problem, he/she would receive a “Cooperation badge”.</td>
</tr>
</tbody>
</table>

All these elements stated in Table 2 had the potential to mitigate student burnout by providing immediate feedback and incentives to motivate and engage students (Sailer & Sailer, 2021). Through teamwork and competition, the incorporation of gamification elements promoted interaction and support among students, increasing their self-confidence and ability to work together (Huang et al., 2019). Points, levels and different types of badges can decrease learning burnout by providing clarity of purpose and a sense of progressive achievement (Mekler et al., 2017).

3.3.2 Gamified Quizzes Sessions

Gamified quizzes were the process of transferring the after-class quizzes from the e-learning platform to the gamified assessment system (Delacruz, 2011). It meant using the existing gamified assessment system to conduct the after-class quizzes. The implementation of gamified assessment involved selecting a suitable system, creating questions, managing class data, and setting evaluation rules. An appropriate gamified assessment system was selected based on factors like functionality and cost. Assessment questions were created, which included a mix of multiple choice and judgement questions and scored out of 100. Class data was effectively managed to ensure the successful implementation of gamified assessment, and evaluation rules were established to ensure fair and accurate assessment.

Gamified quizzes sessions used Quizizz as a tool. Quizizz incorporates two distinct game modes: an assignment mode, which operates at a self-paced tempo, and a classroom mode, which follows a real-time, predetermined pace. The assignment mode was selected for this study to accommodate the learning needs of students after class. Furthermore, various gamification elements were integrated, including points, leaderboards, customizable avatars, auditory cues, countdown timers, and more.
3.4 Data Analysis

Using IBM SPSS 28.0, the quantitative data were analyzed. Cronbach’s alpha was initially determined using the pretest data with the goal of guaranteeing internal consistency for every dimension of the Learning Burnout Inventory-Student Survey in the context of the research. Then, utilizing both pretest and posttest data, descriptive analysis and analysis of covariance were carried out to provide context as well as analyze the overall student learning burnout and each specific dimension between the experimental and control groups. The pretest scores of learning in two groups were considered in this study as factors that may have an impact on the posttest scores but are uncontrollable. Therefore, ANCOVA was used to modify posttest results for variables and remove these undesired consequences, with a confidence interval of 0.95. Participants’ overall learning burnout level in the posttest and each dimension of it were the dependent variables, and the independent variable was “group” (the control group: group=0, the experimental group: group=1). All the basic ANCOVA assumptions were confirmed prior to the analysis. To examine the homogeneity of variance across the two groups, Levene’s test was applied.

4. Results

Cronbach’s alpha was determined using the pretest data to guarantee internal consistency for every dimension of the Learning Burnout Inventory prior to the posttest. All dimensions’ Cronbach’s alphas were more than .7 (α_EX = 0.871, α_CY = 0.939, α_LE = 0.818, α_D = 0.750), indicating that the Learning Burnout Inventory-Student Survey was a trustworthy instrument with internal consistency. The homogeneity of variance was assessed using Levene’s test, and the findings (F_EX = 0.358, p > 0.05; F_CY = 0.469, p > 0.05; F_LE = 6.806, p < 0.05; F_D = 0.485, p > 0.05) indicated that, except from the dimension of low efficacy, there was no variation in the variances across groups. Then after conducting the Kruskal-Wallis test for low efficacy in two groups (P < 0.05), there was also a significant difference in low efficacy. The learning burnout pretest and modified posttest findings are analyzed by descriptive statistics and ANCOVA, as shown in Table 3. The results indicated that the overall learning burnout of the experimental group decreased significantly after the intervention.

Table 3. Descriptive Statistics and ANCOVA Results of Learning Burnout

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Pretest</th>
<th>Posttest</th>
<th>ANCOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Overall Learning Burnout</td>
<td>50.750</td>
<td>11.013</td>
<td>45.684a</td>
</tr>
<tr>
<td>Control Group</td>
<td>52.683</td>
<td>13.346</td>
<td>53.566a</td>
</tr>
<tr>
<td>Emotional Exhaustion (EX)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>17.233</td>
<td>4.681</td>
<td>16.645a</td>
</tr>
<tr>
<td>Control Group</td>
<td>17.033</td>
<td>6.281</td>
<td>17.605a</td>
</tr>
<tr>
<td>Cynicism (CY)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>13.417</td>
<td>4.886</td>
<td>11.204a</td>
</tr>
<tr>
<td>Control Group</td>
<td>13.917</td>
<td>5.419</td>
<td>14.446a</td>
</tr>
<tr>
<td>Low Efficacy (LE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>20.267</td>
<td>5.541</td>
<td>17.835a</td>
</tr>
<tr>
<td>Control Group</td>
<td>21.700</td>
<td>5.564</td>
<td>21.515a</td>
</tr>
</tbody>
</table>

Notes: Experimental Group (n = 60), Control Group (n = 60). The posttest has been modified based on the variables.

To determine whether there were any significant variations in overall learning burnout level and dimensions of learning burnout between the experimental and control groups in the
posttest, repeated measures of ANCOVA were carried out. Overall, students in the experimental group had a lower level of learning burnout than that of the control group, according to the results (Table 3) (F = 13.315, P < .001; \( \eta^2 = .102 \)). In particular, the posttest revealed significant differences between the two groups on cynicism (F = 12.466, P < .001; \( \eta^2 = .096 \)) and low efficacy (F = 17.016, P < .001; \( \eta^2 = .127 \)). The first conclusion is that instructions with gamified assessment in computing curriculum have obviously alleviated the experimental group’s overall learning burnout, especially in terms of cynicism (e.g., students have become less interested in studies since enrollment in the university) and low efficacy (e.g., students can effectively solve the problems that arise in studies). However, there was no significant difference in emotional exhaustion between the experimental and control groups according to the posttest scores. The second conclusion is that gamified assessment did not affect the learning burnout of undergraduate computing students in the emotional exhaustion dimension. Overall, the study found that gamified assessment affects the learning burnout of undergraduate computing students except the emotional exhaustion dimension.

5. Discussion and Conclusion

This study was a pretest-posttest quasi-experimental design to investigate the impact of gamified assessment on learning burnout. The findings of the study demonstrated that gamified assessment had a significant effect on alleviating the learning burnout of undergraduate computing students, particularly regarding cynicism and low efficacy, though no significant difference was manifested in emotional exhaustion. The beneficial effect of gamified assessment in undergraduate education is better recognized. Through instructional practices, it aided teachers in realizing the promising benefits of gamified assessment to alleviate overall and two out of three dimensions of student learning burnout.

The aforementioned results are in line with some research on gamified assessment in education (Dias, 2017; Hassan et al., 2021; Wang & Tahir, 2020; Zainuddin et al., 2020), suggesting that gamified assessment could alleviate burnout in the cynicism and low efficacy. In order to see how gamified assessment affects students’ learning burnout positively or negatively in certain aspect(s), the subsequent stage of this project’s qualitative research has been analyzing the data from student interviews. Academic performance, personality traits, study habits, and family background have all been proposed as potential moderators that could affect students’ learning burnout. Another paper will present and discuss the related findings and conclusions.

The study included several limitations. Firstly, it is still unclear how learning burnout varies in different educational contexts. Future studies are expected to take different subjects’ instructional and students’ individual factors into consideration to draw a wider range of conclusions (Baydas & Cicek, 2019). Additionally, because the intervention only lasted six weeks, its impact on learning burnout may have been rather limited. With longitudinal experiments, further research can be done. Moreover, this study’s sample size of 120 students was small, which might restrict the generalizability of research findings. By using more substantial sample sizes, the reliability can be improved. In addition to the previously mentioned limitations, it is crucial to recognize that language barriers may have caused difficulties for some participants in understanding the scale items. To address this potential issue, future studies could consider involving bilingual participants or conducting separate studies in different language contexts. Furthermore, the data collection mainly relied on self-reporting, which could make results relatively subjective and imprecise. It is possible to conduct extensive study utilizing a variety of tools, such as leveraging machine learning algorithms to assist in classroom observations and involving psychologists in assessing the level of student burnout. Despite the fact that significant differences were found in two out of three dimensions of learning burnout, their effect sizes were minimal (e.g., \( \eta^2_{CY} = 0.096 \)), suggesting that the differences were not substantially significant (Cohen, 2013). There is still a need for more empirical studies investigating the effects of gamified assessment.
This is the first-stage quantitative study investigating the impact of gamified assessment on the learning burnout of undergraduate computing students. The findings of the second-stage qualitative study will explain how gamified assessment affects the learning burnout of undergraduate computing students. The further findings will be reported in another paper. After that, more comprehensive proposals for the incorporation of gamified assessment into tertiary education can be suggested.

Acknowledgements

We would like to express our gratitude to everyone who participated in the study and helped to review the paper.

References


Enhancing Learner Satisfaction in Simulation-Based Learning: The Impact of Learner Characteristics and Expectancy

Steven CK NG*, Li Fern TANb, & Poh Nguk LAUC

aTemasek Polytechnic, Singapore
bTemasek Polytechnic, Singapore
cTemasek Polytechnic, Singapore
*steven_ck_ng@tp.edu.sg

Abstract: Digital game-based learning, facilitated by immersive virtual reality technology, has become integral to modern education. This study investigates learner satisfaction within the context of marketing simulations, exploring the roles of Performance Expectancy (PE) and Effort Expectancy (EE), their interplay with learner characteristics, and their alignment with the Self-Determination Theory (SDT). The study's theoretical framework draws from the Technology Acceptance Model (TAM), Expectancy-Confirmation Theory, and SDT. Through factor analyses and hierarchical regression, the research reveals the significance of Performance and Effort Expectancy in influencing Learner Satisfaction. Notably, Performance Expectancy emerges as a stronger driver than Effort Expectancy. The study also explores the impact of learner characteristics, such as gender and game-playing motivations rooted in SDT's principles of intrinsic motivation, on Learner Satisfaction, identifying Striving as a potential deterrent to satisfaction. Despite limitations like sample specificity and self-reported data, this research provides valuable insights into learner engagement in technology-mediated educational contexts, taking into account both cognitive and motivational aspects. The findings underscore the importance of considering individual motives and characteristics in optimizing educational technology interventions.

Keywords: simulation-based learning, game-based learning, learner satisfaction, learner characteristics, game-playing motives, Self-Determination Theory (SDT), Technology Acceptance Model (TAM), Expectancy-Confirmation Theory, gamified learning environments, educational gamification, higher education

1. Introduction

Digital game-based learning, facilitated by immersive virtual reality technology (Merchant et al., 2014), is integrated into education for engaging and realistic experiences that promote critical thinking (Byun & Joung, 2018; Tsai & Tsai, 2018). Simulations within serious games offer contextualized cognition, skill acquisition, and attitudinal changes (Batko, 2016), fostering interactive and social learning (Buzzard et al., 2011). In STEM fields, simulation-based learning fosters critical thinking, problem-solving, and decision-making skills (Cook et al., 2013; Hegland et al., 2017), while marketing simulations enhance strategic management and decision-making (Hall, 2014; Caruana et al., 2016). Learner satisfaction and user motives are pivotal in simulation effectiveness (Vos & Brennan, 2010; Kolb, 1984), driving the need to explore factors influencing user satisfaction (Caruana et al., 2016).

1.1 Business and Marketing Simulation: Bridging Theory and Practice

Marketing simulations like "Hubro Marketing Simulation (HMS)" and "Markstrat" engage students in decision-making processes (Hall, 2014; Caruana et al., 2016). These simulations offer immediate feedback and objective data on market behaviors (Tompson & Dass, 2000), enabling longitudinal analysis (Kietzmann & Pitt, 2016). Complexity challenges decision-making, with the acceptance and effectiveness of simulations impacting learning outcomes (Vos & Brennan, 2010). User satisfaction is vital, considering (dis)confirmation and specific aspects of the product (Giese & Cote, 2000).
1.2 User Satisfaction as an Outcome Variable

Building on Caruana et al. (2016), this study examines learner satisfaction (Caruana et al., 2016) and explores influencing variables such as gender and age (Caruana et al., 2016). The study integrates the Expectancy-Confirmation Theory and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Caruana et al., 2016). Learner motives, crucial in simulation effectiveness (Deci & Ryan, 1985), are explored alongside user satisfaction, crucial for educational tool optimization (Bartle, 1996; Yee, 2006; Kahn et al., 2015).

1.3 Learner Game-Playing Motivations as a Crucial Part of Learner Characteristics

Learner motives have been identified as critical factors in the context of simulation games, encompassing intrinsic motivation, extrinsic rewards, social interaction, skill development, and personal interest (Deci & Ryan, 1985). Different motives, such as the application of theoretical concepts, hands-on experience, and competition, significantly impact learner engagement, effort, and overall satisfaction with the learning experience. The diversity of student motivation, prior knowledge, and background has been found to have the greatest influence on the effectiveness of teaching interventions (Hattie, 2015). Thus, understanding learner motives is essential for identifying the most effective educational methods.

1.4 Aims of study and research questions

The main focus of this research is to determine the contribution of Performance Expectancy (PE) and Effort Expectancy (EE) on Learner Satisfaction (LS) when the potential impact of Course Performance and Course Type are considered. On the basis of the above discussion, we, therefore, hypothesize that:

1. Hypothesis 1: Performance Expectancy (PE) has a positive influence on Learner Satisfaction with the simulation game.
2. Hypothesis 1a: The effect of Performance Expectancy on Learner Satisfaction is moderated by Learner Characteristics like Gender, Past Course Performance, Course Type, and Learner Game-Playing Motivation
3. Hypothesis 2: Effort Expectancy (EE) has a positive influence on Learner Satisfaction with the simulation game.
4. Hypothesis 2a: The effect of Effort Expectancy on Learner Satisfaction is influenced by Learner Characteristics like Gender, Past Course Performance, Course Type, and Learner Game-Playing Motivation

2. Method

2.1 Participants and simulation deployment

Scholtz and Hughes (2018) emphasized pedagogical alignment, thus meticulously crafting the deployment of the “Hubro Marketing Simulation (HMS)” to optimize connections between Learner Satisfaction (LS), Performance Expectancy (PE), Effort Expectancy (EE), and Learner Motives. Integrated as a group project in a foundational marketing course during October 2022 semester over 12 weeks, teams submitted comprehensive reports detailing marketing choices made over eight rounds, incentivizing competition and sustained decision-making. Eleven tutors, spanning 28 classes, ensured intervention uniformity. Rigorous training aligned with Chaurasia (2017), tutor toolkits, and a practice round mitigated tutor effects for student preparation. Out of 538 students, 227 participated in the survey, including a 37-item questionnaire covering Learner Satisfaction, Performance Expectancy, Effort Expectancy, Learner Motives, and demographic variables. Notably, 66% of respondents were female, reflecting typical gender distribution in Singapore’s business-related diploma courses. Figure 1 illustrates the HMS interface and decision-making guidelines.
2.2 Variables and data analysis

Factor analyses explored the Learner Game-Playing Motivation construct (Jaskari & Henna Syrjälä, 2022) and Learner Satisfaction, Performance Expectancy, and Effort Expectancy constructs. A two-stage hierarchical multiple regression examined hypothesized relationships with 227 respondents, surpassing recommended subject-to-item ratio (Nannally, 1978) and minimum requirement for hierarchical regression (Soper, 2016).

2.2.1 Game-playing motivation construct

Employing established steps (Luomala et al., 2017; Vahlo et al., 2017; Jaskari & Syrjälä, 2022), factor analysis assessed game-playing motivations. Three fixed factors – Immersivity Completing (α = .952), Socializing (α = .874), Striving (α = .909) – explained 78.1% variance (Table 1).

Table 1. Mean and Standard Deviation of Items Together With Pattern Matrix From Factor Analysis for Game-playing Motivation Construct

<table>
<thead>
<tr>
<th>Items</th>
<th>Mean</th>
<th>SD</th>
<th>Loadings</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>When thinking about your own learning, how important do you find the following items?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immersivity Completing</td>
<td>45.74</td>
<td>10.24</td>
<td>0.74</td>
<td>0.95</td>
</tr>
<tr>
<td>1. To advance well in the subject</td>
<td>6.00</td>
<td>1.20</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>2. To be immersed in learning</td>
<td>5.67</td>
<td>1.24</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>3. To feel that you can control your learning</td>
<td>5.79</td>
<td>1.23</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>4. To feel the joy of research in learning</td>
<td>5.59</td>
<td>1.4</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>5. To figure out the idea of marketing and the logic of the subject</td>
<td>5.64</td>
<td>1.31</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>6. To learn the stories and principles related to the topics of the subject</td>
<td>5.54</td>
<td>1.31</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>7. To notice that I have progressed in learning</td>
<td>5.82</td>
<td>1.27</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>8. To set goals and achieve them</td>
<td>5.69</td>
<td>1.28</td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>
2.2.2 Learner Satisfaction, Performance Expectancy, and Effort Expectancy Constructs

Assessed through 7-point Likert scales (1–7), three constructs – LS, PE, EE – underwent factor analysis, aligned with Caruana et al. (2016). The KMO (.96) and Bartlett's test ($\chi^2 = 5812$, p < .001) supported validity. Principal components factor analysis with Oblimin rotation indicated reliability via Cronbach's alpha (.97), surpassing .7 (Nunnally, 1978). (Table 2).

Table 2. Mean and Standard Deviation of Items Together With Pattern Matrix From Factor Analysis for Learner Satisfaction, Performance Expectancy, and Effort Expectancy constructs

<table>
<thead>
<tr>
<th>Items</th>
<th>Mean</th>
<th>SD</th>
<th>Loadings</th>
<th>Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learner Satisfaction (LS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The use of the simulation game in the subject was one of the best experiences I could have had.</td>
<td>5.16</td>
<td>1.37</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>I am satisfied with having used the simulation game in the subject.</td>
<td>5.31</td>
<td>1.44</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>The choice of the simulation game for the subject was a wise one.</td>
<td>5.28</td>
<td>1.52</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>I am certain it was the right thing that the simulation game was included as an integral part of the subject.</td>
<td>5.11</td>
<td>1.42</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>Playing the simulation game engaged me more in the course than a regular teaching method does.</td>
<td>5.18</td>
<td>1.58</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>I will recommend the use of simulation game to learn and apply marketing principles and theories.</td>
<td>5.22</td>
<td>1.67</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td><strong>Performance Expectancy (PE)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using the simulation game enabled me to accomplish tasks in the subject more quickly.</td>
<td>5.06</td>
<td>1.52</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Using the simulation game improved my study performance in this subject.</td>
<td>4.82</td>
<td>1.56</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Using the simulation game improved my understanding of marketing.</td>
<td>5.30</td>
<td>1.48</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Using the simulation game enhanced my effectiveness in my study of marketing.</td>
<td>5.08</td>
<td>1.52</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Using the simulation game made it easier to develop marketing competencies.</td>
<td>5.21</td>
<td>1.41</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>I found the simulation game useful for addressing marketing-related issues.</td>
<td>5.26</td>
<td>1.44</td>
<td>0.94</td>
<td></td>
</tr>
</tbody>
</table>
Effort Expectancy (EE) | 29.75 | 9.28 | 0.97
---|---|---|---
Learning to use the simulation game was easy for me. | 4.89 | 1.61 | 0.98
It was easy to get the simulation game to do what I wanted it to do. | 4.94 | 1.56 | 0.88
Interfacing with the simulation game was clear and understandable. | 4.99 | 1.55 | 0.78
I found the simulation game flexible to interact with. | 5.07 | 1.44 | 0.74
It was easy for me to become skillful at using the simulation game. | 4.82 | 1.57 | 0.77
I found the simulation game easy to use. | 5.04 | 1.55 | 0.86

Note. Extraction method: Principal components analysis. Rotation method: Varimax. Loading coefficients from the factor analysis below .5 are not shown.

2.3 Relationship between Variables

Following Caruana et al. (2016), a two-stage hierarchical regression method was applied. In Stage 2, two independent variables were added beyond Stage 1 variables, ensuring normality and linearity assumptions. Harman's single-factor test revealed no significant common method bias (Podsakoff et al., 2003).

Stage 1 of the hierarchical regression involved Gender, Course Type, Past GPA Attainment, and game-playing motivations (Immersivity Completing, Socializing, Striving) as independent variables predicting Learner Satisfaction. The results (Table 3) explained 41% variance. Significant predictors were Gender (β = 0.24, p < .05), Immersivity Completing (β = 0.53, p < .001), Socializing (β = 0.30, p < .001), and Striving (β = -0.27, p < .001). (Table 3)

Table 3. Results From the Two-Stage Hierarchical Multiple Regression (N = 227).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Stage 1</th>
<th></th>
<th>Stage 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std. β</td>
<td>t</td>
<td>sr²</td>
<td>Std. β</td>
</tr>
<tr>
<td>Gender</td>
<td>0.24</td>
<td>2.16*</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Course Type</td>
<td>0.01</td>
<td>0.28</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>Past GPA Attainment</td>
<td>0.15</td>
<td>1.48</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Immersivity Completing</td>
<td>0.53</td>
<td>6.76***</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>Socializing</td>
<td>0.30</td>
<td>3.36***</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>Striving</td>
<td>0.27</td>
<td>-3.67***</td>
<td>0.04</td>
<td>-0.09</td>
</tr>
<tr>
<td>Performance Expectancy</td>
<td>0.52</td>
<td>10.67***</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Effort Expectancy</td>
<td>0.41</td>
<td>8.98***</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0.66</td>
<td>9.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.43</td>
<td>8.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjusted R2</td>
<td>0.41</td>
<td>8.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change R2</td>
<td>0.41</td>
<td>8.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p<.05 **p<.01 ***p<.001

Aligned with Caruana et al. (2016), a robust Performance Expectancy-Effort Expectancy correlation drove regression score computation via exploratory factor analysis. The model variance explained 84%, F(8, 218) = 148, p < .001. Performance Expectancy, Effort Expectancy, and Striving contributed 14% to Learner Satisfaction variance post adjustment for Gender, Past GPA Attainment, and Course Type. In the final model, Performance Expectancy (β = 0.52, p < .001) surpassed Effort Expectancy (β = 0.41, p < .001), and Striving (β = -0.09, p < .05) was also notable (Table 3).
3. Results and Discussion

The analysis of research hypotheses and their corresponding outcomes, as presented in Table 4, furnishes a comprehensive comprehension of the intricate interplay between Performance and Effort Expectancy and their influence on Learner Satisfaction. The meticulous assessment revealed a substantial explanatory capacity of 84.4% concerning Learner Satisfaction. Importantly, both Performance and Effort Expectancy emerged as significant drivers, aligning harmoniously with established research within the Technology Acceptance Model (TAM) framework (Caruana et al., 2016; Davis, 1989; Shen & Eder, 2009; Wu & Gao, 2011). Noteworthy is the prominence of Performance Expectancy, surpassing Effort Expectancy, thereby accentuating its pivotal role.

An intriguing facet of the findings pertains to the partial validation of Striving as a game-playing motivation and its impact on the model's outcomes, diverging from Jaskari and Syrjälä’s (2022) findings. Methodological distinctions are likely contributors; our study examined individual motives, whereas earlier research concentrated on aggregated clusters. The counterintuitive effect of Striving may arise from learners perceiving a curtailed sense of control, particularly in competitive scenarios where their academic performance is at stake, an aspect particularly pertinent in the Singaporean context where grades hold paramount importance for university admission. In response, our team aimed to experiment with game elements like awarding "Most Valuable Player" among high achievers within the teams to mitigate the potential negative impact of Striving.

Additionally, Gender emerged as a determinant, notably evident in Stage 2, where female learners exhibited relatively lower satisfaction compared to males, aligning with technology adoption and simulation-based educational games literature (Caruana et al., 2016; Venkatesh et al., 2003; Fenwick & Neal, 2001), underscoring the importance of addressing gender-based disparities in educational technology interventions. To summarize, the systematic exploration of Performance and Effort Expectancy, coupled with the nuanced influence of game-playing motivation and gender, imparts valuable insights into the multifaceted dynamics underpinning Learner Satisfaction. These findings contribute to a broader comprehension of learner engagement within technology-mediated educational contexts, thereby offering implications for instructional design and suggesting avenues for further research in the field.

Table 4. Summary of Hypotheses and Conclusions on Performance and Effort Expectancy Effects on Learner Satisfaction

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Accepted:</strong> Strongly supported. Results from a two-stage hierarchical regression confirm a positive correlation between Performance Expectancy (PE) and Learner Satisfaction, in alignment with established models (TAM) (Caruana et al., 2016; Davis, 1989).</td>
</tr>
<tr>
<td>1a</td>
<td><strong>Accept with Consideration:</strong> The impact of Performance Expectancy on Learner Satisfaction is affirmed, while the significance of Gender becomes apparent. Female participants exhibit lower satisfaction levels, indicating a moderated effect of Performance Expectancy by Gender. Further investigation into game-playing motivations and course-related attributes is crucial for a nuanced comprehension.</td>
</tr>
<tr>
<td>2</td>
<td><strong>Accepted:</strong> Strongly supported. Substantial evidence confirms that Effort Expectancy (EE) is a significant predictor of Learner Satisfaction, aligning with established theoretical frameworks (TAM).</td>
</tr>
<tr>
<td>2a</td>
<td><strong>Accept with Reservation:</strong> Partial support The influence of Effort Expectancy on Learner Satisfaction is validated; however, the interplay between Learner Characteristics such as Gender, Past Course Performance, Course Type, and Learner Game-Playing Motivation warrants deeper exploration.</td>
</tr>
</tbody>
</table>
4. Limitations and Directions for Further Research

This study responds to the call for deeper cross-cultural investigations, as advocated by Caruana et al. (2016) and Jaskari & Henna Syrjälä (2022), but is confined to a specific polytechnic business school in Singapore, limiting generalizability across diverse settings. To enhance applicability, future research should explore various educational contexts.

The reliance on self-reported data introduces potential bias and subjective interpretations. Incorporating objective measures or observational data would enhance result reliability. The cross-sectional design precludes causal inferences and temporal insights; adopting longitudinal or experimental approaches could address these limitations.

While this study focuses on Performance and Effort Expectancy, other variables like self-efficacy and engagement contribute to Learner Satisfaction. Future studies should consider broader variable inclusion to refine understanding. In conclusion, while shedding light on Performance and Effort Expectancy's impact, this study acknowledges contextual boundaries, self-report reliance, and the need for variable expansion. Overcoming these limitations would refine comprehension of technology-mediated education engagement.

Acknowledgements

We thank Mr Matthew Chua and the team of Marketing Fundamentals at Temasek Polytechnic, School of Business, for their assistance in launching the simulation unit.

References


Exploring the Impact on Student Reading Preferences in Gamified Reading Portfolio

Hsiao-Tung YANG1*, Chang-Yen LIAO2, Ciao-Min SYU1 & Tak-Wai CHAN1

1Department of Network Learning Technology, National Central University, Taiwan
2Department of Hakka Language and Social Sciences, National Central University, Taiwan
*jessica4625@gmail.com

Abstract: Extensive reading enhances horizons, deepens understanding, and nurtures critical thinking. This study aims to gamify reading goals, turning teacher-recommended books into engaging game levels. Books are categorized and distributed by genre and depth. And we will design achievement medals, allowing students to earn medals upon meeting specific criteria. The goal is to boost reading interest, fostering a passion for learning through accomplishments. This research addresses selective reading, encouraging exploration of varied genres while nurturing a genuine love for reading.

Keywords: Gamification, Achievement Goals, Reading Interest, Extensive Reading

1. Introduction

Reading initiatives like the "morning reading program" and "reading courses" are well-established in Taiwan. Since 2009, our team has researched Modeled Sustained Silent Reading (MSSR) and introduced the "morning reading campaign" to schools. Promoting MSSR actively nurtures a reading habit and enjoyment. However, as students mature, a challenge arises: they develop "reading preferences", sticking to specific genres and avoiding diverse books. Some also resist advancing to complex levels, staying in their "comfort zone" and limiting knowledge expansion through reading.

Therefore, the "Gamified Reading Performance System" designed by this institute incorporates a list of books recommended by several language teachers for grades 1 to 6 into 100 levels in order to help students read different types and depths of books by completing different levels, so that students can choose the level that suits their reading according to their own interests or abilities, and then record the books using the system after finishing the reading. After reading the books, use the system to record the books. After reading the books and completing the levels, students can receive corresponding rewards to keep them motivated to continue the challenge.

This study investigates how the system can help teachers and students in schools to make reading activities more attractive and interesting. Through the design of this system, reading is no longer just "reading a book", but can become an activity that can interact with others and express oneself, thus increasing the interest and motivation of reading, which is the goal of this study.

2. Literature Review

2.1 Gamification

Gamification involves integrating game elements into non-game contexts to enhance engagement and motivation. By employing techniques like points, badges, and leaderboards, it aims to make tasks more enjoyable. (Deterding et al., 2011). In education, gamification keeps learners immersed and time-conscious, impacting their flow experience (Mak, Wang & Chu, 2019). Effective design involves tasks aligned with learners' interests and autonomous goal-setting (Deci & Ryan, 2000).
2.2 Goal Setting

The theory of goal setting, originating from Locke and Latham's seminal work (2002), underscores the potency of establishing distinct and demanding objectives to amplify motivation, concentration, and competence. The lucidity and challenge of goals significantly influence their accomplishment. Well-defined targets contribute to positive sentiments, contentment, and self-efficacy. A clear self-assessment process further bolsters self-efficacy. Proficient goal setting, characterized by quantifiable milestones and time-sensitive benchmarks, not only bolsters productivity but also fosters personal development and triumph across diverse facets of life (Locke et al., 1981).

3. System Design

Our research team introduced a reading system to schools in Taiwan in tandem with reading programs. This system enables students to record their reading progress, motivations for book choices, and recommendations. Notably, many students display reading habits but lack long-term goals. Thus, we aim to employ Goal Setting and Gamified Learning theories, integrating an "Achievement Medal Mechanism" and "Data Visualization." This aids students in pursuing personalized reading interests, guided by level coordination. A "Reading Dashboard" (Figure 1) monitors indicators like level progress and book categories. The main display features 100 levels, each containing varying books. Passing a level requires reading a set number of books, marked by changing button color and "rockets" earned, encouraging challenge and sustained reading motivation.

Upon entering a level (Figure 2, you can access basic information about the books, including titles, categories, and depth. Three book categories (story fiction, humanities and society, science and technology) with four sub-categories each (e.g., warm and inspirational, adventure and reasoning) are featured. Book depth is rated by stars: 1 star for picture books, 2 stars for bridge books, 3 stars for beginner texts, 4 stars for intermediate texts, and 5 stars for advanced texts. Each level's top displays the total "stars" earned by "read books." A checkmark symbolizes completed books, while an unread lock indicates locked books. By equitably distributing various book types and depths, students are exposed to a broader reading spectrum, fostering a balanced reading habit and preventing preferences.

The achievement medal collection (Figure 3 presents medals based on book types and depths, categorized as bronze, silver, and gold. These medals can be sorted by type, and a click on "Your Achievements" reveals all earned medals. By earning these, students enhance self-confidence, fostering diverse reading preferences and expanding reading horizons.

Finally, Figure 4 resents the "Reading Leaderboard" showcasing top students in each index, encouraging competition and improvement. "My Reading Performance" displays weekly detailed reading status, aiding students in self-review, comparison, and refining reading plans.
4. Result

This study involves 30 students from a primary school in Taiwan, specifically the second and third grades. They engage in daily Modeled Sustained Silent Reading (MSSR) activities during an eight-week experiment. The system was developed in alignment with the school's reading program after understanding the requirements from teachers and students. Using the Design-Based Research (DBR) approach, bi-weekly interviews were conducted after implementing the system to collect feedback and make gradual adjustments. Log files were also gathered for user behavior analysis.

Two interview rounds focused on students' system usage, feedback, and its impact on reading behavior. Initial findings indicate that about 40% of participants explore unfamiliar book genres through the level arrangement, while around 30% intend to read more diverse books to complete additional levels. These preliminary interviews highlighted persisting "reading preferences" among students. Therefore, we plan to enhance the system further. For example, we aim to introduce personalized prompts on the dashboard, offering individualized adjustments and suggestions based on reading progress. Future research will include alternative methods to assess comprehension. Beyond setting reading quantity goals, we contemplate integrating activities like recording impressive content, sharing thoughts through videos after completing a book, and recommending good books to peers. These could serve as level completion criteria or achievement objectives, enhancing reading motivation. The incorporation of diverse activities to amplify enjoyment and challenges in reading could provide insights into their motivational impact on students.

Acknowledgements

This study was funded by the National Science and Technology Council of Taiwan (NSTC 112-2410-H-008-020-MY3) and by Research Center for Science and Technology for Learning, National Central University, Taiwan.

References


Game-based learning provides a promising methodology for teaching skills such as teamwork (Martín-Hernández et al., 2021; Riivari et al., 2021; Syynimaa et al., 2022). Studies of human teamwork have identified various team processes that underlie good joint task performance (e.g., Marks et al., 2001). The multiagent research community has operationalized many of these findings into domain-independent algorithms that separate task-specific knowledge from reusable task-independent teamwork knowledge (e.g., Tambe, 1997). These frameworks were successfully deployed in facilitating teamwork in real-world human teams when supplied with task-specific knowledge (e.g., Chalupsky et al., 2001).

We present here an agent that focuses exclusively on teamwork-level variables in deciding what interventions to use in coaching a human team in a game-based search-and-rescue environment in Minecraft (Corral et al., 2021). The game is an urban search and rescue (USAR) task that involves clearing and avoiding hazards while rescuing victims of a disaster. The task is performed by distributed teams of three participants, possibly with a human or agent as a coach. The coaches are not acting team members, but they are able to communicate, via text chat, with the human team members, who themselves can directly engage each other in high-volume, unconstrained voice communication.

Our agent does not directly observe the game environment or the people in it, but instead relies on input from analytic components (ACs) (developed by other research teams) that process environmental information and output only teamwork-relevant measures. Our agent models teamwork variables and updates its beliefs over them using PsychSim, which provides reusable AI technology for agents capable of populating game environments (Pynadath & Marsella, 2005). PsychSim integrates two AI technologies: recursive models (Gmytrasiewicz & Durfee, 1995) and decision-theoretic reasoning in the form of POMDPs (Kaelbling et al., 1998). PsychSim has been used to build interactive training games for a variety of soft skills, such as negotiation (Kim et al., 2009) and avoiding risky behavior under social pressure (Miller et al., 2011). For the current search-and-rescue task, we built an agent that assesses the team it is currently observing and chooses interventions to best assist them. Details on our agent and some objective measures of its purposes are available in an earlier publication (Pynadath et al., 2023). In this paper, we look deeper into some specific observations of human behavior when interacting with our agent in the game.

Our agent used team-process variables (Marks et al., 2001) whose measures were recently validated (Mathieu et al., 2020), but only those relevant to the game environment:

- **Affect management**: "activities that foster emotional balance, togetherness, and effective coping with stressful demands and frustration."
Coordination: “the process of synchronizing or aligning the members’ actions.”

Motivating and confidence building: “activities that develop and maintain members’ motivation and confidence while working toward team goals.”

Systems monitoring: “activities such as tracking team resources (e.g., money) and factors in the team environment (e.g., inventories) to ensure that the team has what it needs to accomplish its goals and objectives.”

Team monitoring: “members assisting others in the performance of their tasks (by providing feedback or coaching or assisting with the task itself)." None of these processes are directly observable. We thus model them as hidden variables that our agent seeks to improve through the following candidate interventions:

- **Reflection:** Between trials, the agent prompts the team with a reminder about a situation in the first trial where a player was stuck on a threat plate for an inordinate amount of time.

- **Cheerleading:** The agent congratulates a player on successfully achieving a goal (more specifically, moving a victim to a triage area).

- **Report drop:** The agent reports on a noteworthy lack of performance by one player (more specifically, failure to respond to outstanding requests by a teammate).

- **Notify phase (early):** The agent reminds the team that it is still early in the game, so exploration should be valued more.

- **Notify phase (late):** The agent reminds the team that it is getting late in the game, so exploration should be valued less.

- **Remind of best practices:** The agent suggests that someone help a player who has a number of outstanding, but unaddressed, requests.

- **Prompt activity:** The agent asks about any possible issues upon observing that the team has been predominantly idle for a period of time.

Data for thirteen teams were collected at Arizona State University. One immediate question was whether the teams even noticed our agent’s interventions. For one team, five of eight interventions were verbally acknowledged; for two others, none were. For all other teams, one or two interventions were explicitly acknowledged. Breaking down these counts by intervention type, none of the cheerleading interventions were acknowledged. The two notify phase interventions were the most commonly acknowledged (14 of 38), while the various request/activity based interventions were less frequently acknowledged (8 of 38).

The post-trial survey asked a few questions about the participants’ opinions of their coach. Not surprisingly, the human coaches received much higher trust scores than the agent coaches. All of the agent coaches were distrusted, but ours was distrusted the least.

Beyond these aggregate results, we found it necessary to examine the actual behaviors of individual teams, in addition to listening to the dialogue within those teams. The relatively small number of teams makes this in-depth examination feasible, although it also potentially limits the generalizability of the observations made.

One recurring issue was a disconnect between what our agent perceived as requests vs. what the team members actually intended as requests. Participants could mark blocks in Minecraft (e.g., for victims, for obstructed rooms), making an implicit request (e.g., to evacuate/triage the victim inside, to clear the obstruction). However, marking these blocks was sometimes used as implicit communication or even just a mnemonic device, not a request. This difference of perspective led to occasional confusion when our agent intervened to stimulate satisfaction of “requests”. For example, in one trial, our agent performed a remind of best practices intervention regarding Red's requests, but Red then expressed no awareness of any requests when asked by another team member to clarify.

As a more positive example, there was another team where our agent also chose remind of best practices for Red and where a team member also asked for clarification. However, in this case, Red responded by explicitly identifying a victim, its type, and its location. The other team member then saw this victim and triaged it. Our agent's intervention thus triggered the desired shared situational awareness and subsequent task completion.

A more ambiguous example starts off just as these first two, with a team member asking for clarification of Red's requests after an agent intervention. In this case, Red admits
to accidentally forgetting to pick up a marker. Despite our agent’s mistaken reaction to a nonexistent request, the communication prompted by its intervention preempted the misunderstanding that would have occurred if the discussion had not taken place.

As another example of the evaluation challenge, our agent performed the late version of its “notify phase” interventions while the team was in the middle of discussing its strategy. There was a pause while the speaker considered the agent's message, but then the team resumed discussing and executing its current plan, which was in conflict with our agent’s recommendation. The team was clearly not complying, but it is not clear whether they were better or worse off as a result. Thus, while the lack of compliance is objectively measurable, any wider-scope learning outcomes are harder to evaluate.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-20-1-0011 and by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Any opinions, findings, or recommendations do not necessarily reflect the views of DARPA or the US Government.

References


TELL
Using Flipped Classroom Approach to Integrate SRL Instruction in Classical Chinese Reading Instruction: Insights from the First-Year Results

Kit Ling LAUa
aThe Chinese University of Hong Kong, Hong Kong
*dinkylau@cuhk.edu.hk

Abstract: This study adopted a flipped classroom (FC) approach to combine the instructional principles of self-regulated learning (SRL) and eLearning activities in a two-year intervention program to facilitate Hong Kong students' learning of classical Chinese (CC). Using a quasi-experimental design, students from three Hong Kong secondary schools were assigned to three different treatment groups: SRL instruction plus out-of-class eLearning, SRL instruction only, and control. Pre- and post-treatment measures of CC reading comprehension tests and student questionnaires were used to evaluate the effectiveness of the program on enhancing students' strategy use, motivation, and performance in CC reading. This paper reports the findings from the quantitative data collected in the first-year of the study. The findings indicated that while all treatment groups showed significant improvements in the CC reading post-test, only students in the two experimental groups significantly increased their use of CC reading strategies, self-efficacy, and intrinsic motivation. The findings offer promising preliminary evidence supporting the viability and advantages of integrating SRL instruction and FC as an effective instructional approach to enhance student learning within a conventional academic domain.

Keywords: Classical Chinese reading, eLearning, flipped classroom, motivation, self-regulated learning instruction

1. Introduction

Self-regulated learning (SRL), involving metacognition, motivation, and strategic actions, has been widely advocated as an effective form of learning (Winne & Perry, 2000; Zimmerman, 2000). Schunk and Zimmerman (1997) proposed four developmental levels through which learners progress from other-regulation to self-regulation. At the observational level, learners observe the key aspects of strategies by watching a proficient model. Moving to the emulation level, learners attempt to apply the observed strategies by imitating the patterns while being supervised by the model. In the self-controlled level, learners gain the ability to use the strategies independently. Finally, in the self-regulation level, learners can systematically regulate their behaviors and performance in response to varying conditions. Based on previous SRL research, the author proposed a "TSAE" framework (Lau, 2013; 2020) to guide teachers in fostering students' progression towards higher levels of SRL. The framework consists of four major instructional principles: Task nature (T), where teachers provide direct instruction on subject knowledge and learning strategies while designing open tasks to facilitate knowledge transfer. Teacher support (S), where various scaffolding techniques are employed to assist students in executing self-regulatory skills. Student autonomy (A), where student-directed activities are designed to enable practice of self-regulatory skills. Evaluation practices (E), which involve students in self- and peer-evaluation to enhance metacognitive knowledge and skills (Cousins et al., 2022; Dignath et al., 2008; Dignath & Veenman, 2021; Jayawardena et al., 2019; Lombaerts et al., 2009; Mohammadi et al., 2020; Neitzel & Connor, 2017; Perry & Rahim, 2011).
Using the TSAE framework, the author developed an intervention program to improve Hong Kong students’ classical Chinese (CC) reading, traditionally dominated by teacher-centered instruction (Lau, 2020). While the findings of that study indicated that SRL instruction was more effective than the traditional teacher-centered instruction in enhancing students’ prior CC knowledge and comprehension, no significant changes in their strategy use and motivation were found. Given the potential limitations of the intervention effects due to students’ weak foundation in CC reading and the short duration of the previous study, this study aimed to expand upon the author’s 2020 study by incorporating a flipped classroom (FC) approach to add an eLearning component into SRL instruction and extending the intervention duration to two years. FC is a widely adopted blended pedagogical approach that comprises two major components: in-class instruction and out-of-class eLearning. By having students study content before class through online learning to free in-class time for more interactive and higher-level learning activities, FC reverses the traditional teacher-centered classroom into student-centered learning (Berrett, 2012; Fulton, 2012; Strelan et al., 2020). Compared with solely online learning, FC put more emphasis on teachers’ role as a facilitator in face-to-face teaching to enable students to perform SRL (Öztürk & Çakıroğlu, 2021; Sletten, 2017), which is consistent with the Principle S of SRL instruction. At the same time, the out-of-class eLearning component of FC allows students to practice various SRL skills independently, which is better than the traditional classroom instruction in implementing the Principle A of SRL instruction (Al Mulhim, 2021; Rasheed et al., 2020).

Specifically, the study aimed to examine (1) whether SRL instruction was more effective than the traditional teacher-centered instruction on enhancing students’ strategy use, motivation, and performance in CC reading and (2) whether the inclusion of out-of-class eLearning activities produced more positive effects on enhancing students’ learning of CC reading than when only SRL instruction was used.

Compared with many short-term SRL studies, the study adopted a longitudinal design to guide students’ development of SRL through the four developmental levels. Moreover, considering that CC reading is very difficult for Chinese students and has long been dominated by teacher-centered instruction, it was particularly interested for the study to verify whether applying SRL instruction and FC in CC reading instruction could achieve similar positive effects as in other subject areas. The findings should provide valuable insights into the applicability of SRL instruction and FC in traditional subject areas.

2. Methodology

2.1 Participants

The study adopted a quasi-experimental design. A total of 352 Grade 8 students (172 boys; 179 girls; 1 unreported) between the ages of 12 and 14 years (mean = 12.94 years, SD = .63) from three secondary schools in Hong Kong participated in the study on a voluntary basis. To ensure the comparability of students from different schools, all the three schools mainly admit moderate-achieving students, use the same Chinese language textbook, and have a similar number of Chinese language lessons each week. Students from each school were assigned to one of the three treatment groups: SRL instruction plus out-of-class eLearning (EG-I+E), SRL instruction only (EG-I), and control (CG). Students in EG-I+E and EG-I had the same number of Chinese language lessons, CC reading materials, and in-class SRL instruction. While students in EG-I+E were assigned out-of-class eLearning activities using a FC approach, EG-I students received post-class paper-based SRL tasks. Teachers of the CG used traditional teacher-centered instruction and assigned the same CC texts to their students in the form of traditional comprehension exercises.

2.2 Design of the intervention program

The design of the intervention program was based on the TSAE framework and the four-stage SRL development model. The intervention program comprised four phases. One CC
reading module was designed for each phase and implemented in one of the semesters of the two academic years. Each module consists of twenty 40-minutes lessons. Different CC texts were organized based on a humanistic theme of ancient Chinese culture and various types of CC knowledge and reading strategies were included in each module to enhance students’ CC reading ability. Interesting materials and interactive in-class learning activities were designed to enhance student motivation. Pre- and post-class eLearning activities were provided for students to make pre-class preparation and practice the knowledge and strategies in new tasks. The importance of the TSAE instructional principles was adjusted according to students’ developmental levels of self-regulatory skills. The detailed description of the instructional design of each phase can be found in Lau (2021).

2.3 Measures

2.3.1 CC Reading Comprehension Tests

Two CC reading comprehension tests were designed to assess students’ CC reading ability in each academic year. Each test comprised one narrative and one argumentative text. Five “word interpretation” questions and nine “text comprehension” questions were designed for each type of text to assess students’ word- and text-level reading performance. The internal consistency reliabilities of the pre- and post-reading test were .66 and .69, respectively.

2.3.2 Student Questionnaires

Two validated student questionnaires were adopted from the author’s studies (Lau, 2019; 2020) to assess students’ strategy use and motivation in CC reading. The CC reading strategy questionnaire comprised two subscales, “text comprehension strategies” and “word interpretation strategies” subscales, with 26 items to measure students’ use of different strategies during their reading of CC texts. The internal consistency reliabilities of these two subscales in pre- and post-test measures ranged from .85 to .91. The CC reading motivation questionnaire comprised three six-item subscales measuring students’ self-efficacy, intrinsic motivation, and extrinsic motivation in CC reading. The internal consistency reliabilities of these three subscales in pre- and post-test measures ranged from .86 to .92. All items in the two questionnaires were rated on a 5-point Likert scale.

2.4 Procedures

The instructional design of the program and the CC reading tests were reviewed by two scholars and all EG teachers. The reading tests were piloted in a school with a similar background to the experimental schools and refinements were made based on reviewers’ comments and the pilot study’s results. The intervention started from the academic year of 2021/2022 and ended at the academic year of 2022/2023. Teacher workshops and meetings were held for EG teachers to support their preparation and evaluation of implementing the experimental materials. All pre-tests were administered in September 2021 and the first and second wave of post-tests were administered in July 2022 and June 2023, respectively.

3. Results of the First Year

4.1 Student changes on CC Reading performance

Repeated measures ANOVA using time as the within-subject factor and treatment as the between-group factor were performed to compare changes among different treatment groups. The findings indicated a significant time effect on reading comprehension test scores (F = 99.33, p < .001, Partial η² = .227) but no significant time × group interaction effects were found (F = 1.53, p = .218, Partial η² = .009). The results of follow-up paired samples t-tests indicated that all treatment groups significantly increased their reading comprehension test scores at the end of the first academic year (EG-I+E: t = 5.18, p < .001; EG-I: t = 4.47, p < .001; CG: t = 7.45, p < .001).
4.2 Student changes on Strategy Use in CC Reading
Repeated measures ANOVA indicated significant time effects in both word-level strategies (F = 52.70, p < .001, Partial \( \eta^2 = .141 \)) and text-level strategies (F = 18.41, p < .001, Partial \( \eta^2 = .056 \)). A significant time \times group interaction effect was found in word-level strategies (F = 12.76, p < .001, Partial \( \eta^2 = .074 \)). The results of paired samples t-tests indicated that while students in the two EGs had similar and significant increases in their use of word-level strategies (EG-I+E: t = 5.34, p < .001; EG-I: t = 6.52, p < .001), CG students did not have any significant change in the reading strategy post-test (t = .19, p = .852).

4.3 Student changes on Motivation in CC Reading
Repeated measures ANOVA indicated significant time effects in self-efficacy (F = 19.35, p < .001, Partial \( \eta^2 = .055 \)) and intrinsic motivation (F = 11.56, p < .01, Partial \( \eta^2 = .034 \)). A significant time \times group interaction effect was also found in both types of motivation (self-efficacy: F = 52.70, p < .001, Partial \( \eta^2 = .141 \); intrinsic motivation: F = 52.70, p < .001, Partial \( \eta^2 = .141 \)). The results of paired samples t-tests indicated that while students in the two EGs showed similar and significant improvement in their self-efficacy (EG-I+E: t = 3.54, p < .01; EG-I: t = 3.80, p < .001) and intrinsic motivation (EG-I+E: t = 2.59, p < .05; EG-I: t = 2.91, p < .01), CG students did not have any significant motivational change in the post-test (self-efficacy: t = .16, p = .113; intrinsic motivation: t = 29, p = .114).

4. Discussion and Conclusion
The significant improvement of EG students’ and CC reading performance and motivation was consistent with previous SRL studies (e.g., Bai et al., 2021; Cousins et al., 2022; Mohammadi et al., 2020; Otto & Kistner, 2017) and provided empirical evidence for expanding the SRL research to the context of CC learning, indicating that SRL instruction is also a feasible and effective method for enhancing students’ CC learning. Since students’ major difficulties in CC reading are at the lexical level (Chen & Chen, 2020; Lau, 2019), the first two modules of the intervention program focused on the teaching of word interpretation strategies to provide concrete ways for students to understand unfamiliar CC words. The significant increase in EG students’ use of word interpretation strategies together with their better reading performance support the importance of direct strategy instruction in SRL instruction (Dignath & Büttner, 2018; Greene et al., 2015). Since most students regard CC reading as very difficult, they usually have poor self-efficacy in CC reading (Lau, 2019; Chi & Chiu, 2015). The increases of EG students’ self-efficacy, strategy use, and reading performance are consistent with the view that students’ self-efficacy can be enhanced through strategy learning and successful learning experiences (Schunk & Ertmer, 2000). Moreover, the significant improvement in EG students’ intrinsic motivation supports that interesting and authentic instructional tasks and a high degree of autonomy can enhance students’ intrinsic interest in learning (Dignath et al., 2008; Perry et al., 2002).

Contrasting to the view that teacher-centered instruction is not effective in enhancing students’ CC reading performance (Lau, 2019; Liu, 2020), this study found that CG students also demonstrated significant improvement in the CC reading post-test. As all treatment groups read a same number of CC texts during the study, this finding suggests that while EG students improved their CC reading performance through consciously applying the knowledge and strategies they learned in the intervention program, CG students might improve their CC reading ability through having more practices in doing CC reading exercises. However, it is noteworthy that although CG students improved their performance in the reading comprehension post-test, their self-efficacy and intrinsic motivation was not enhanced because they only learned passively under the control of their teachers.

Both EG-I+E and EG-I had similar improvements in all post-test measures, suggesting that adding the eLearning component in SRL instruction did not achieve additional positive effect on student learning. The major advantage of eLearning in promoting SRL is providing a flexible and autonomous environment for students to practice
their SRL skills (Blau & Shamir-Inbal, 2017; So et al., 2019). As student autonomy is also emphasized in the Principle A of SRL instruction (Perry & Rahim, 2011; van Grinsven & Tillema, 2006), EG-I students might also perceive a high degree of autonomy when participating in the student-centered in-class activities. Given that the in-class instruction for both EGs was identical, it is possible that similar positive effects on students’ learning were achieved in both groups.

In conclusion, the findings of the first-year implementation of the intervention program provide initial support for the effectiveness of using FC to integrate SRL instruction and eLearning to facilitate student learning in the context of CC reading. Despite CC reading being considered a challenging and teacher-centered subject, this study’s intervention program design offers concrete ways for teachers to guide students in the gradual development of their SRL skills for CC reading. However, the similar improvements revealed between the two EGs did not support the prediction that adding the eLearning component should facilitate student SRL better than the traditional face-to-face instruction. The possible benefits of using FC to increase the effectiveness of SRL instruction should be explored further using the results of the second year when the degree of student autonomy and the difficulty level of the learning content is increased in the last two phrases. Finally, despite the similar backgrounds of the groups, controlling all confounding factors was challenging due to practical constraints in implementing a longitudinal intervention program within the students’ regular classes. To ensure better comparability between different treatment groups, future studies should consider measures such as assigning the same teachers to teach all treatment groups and using the pre-tests results to screen participants into different groups.

Acknowledgements

This study was supported by the Hong Kong Research Grants Council (RGC 14608820).

References


Relations between Instructional Factors and Student Acceptance of Flipped Learning in Chinese Language Learning

Kit Ling LAU\textsuperscript{a} & Quan QIAN\textsuperscript{b}
\textsuperscript{a}The Chinese University of Hong Kong, Hong Kong
\textsuperscript{b}Beijing Normal University, Zhuhai
\textsuperscript{*dinkylau@cuhk.edu.hk}

Abstract: This study investigated the relations between instructional factors and students' acceptance and use of flipped learning (FL) in the context of L1 Chinese language learning. A total of 2160 students from ten secondary schools in Hong Kong filled out two questionnaires measuring their perception of the instructional design and implementation of FL in their Chinese language classes and their acceptance of and actual participation in FL activities. Findings of the descriptive analysis indicated that from the perception of students, the design and implementation of FL in Chinese language classes generally adhered to the instructional principles of FL. Students also showed a moderately high level of acceptance of FL and a moderate level of participation in FL activities. The results of structural equation modeling indicated that both the quality of in-class and out-of-class eLearning activities had significant and positive effects on students' perceived usefulness, ease of use, and enjoyment of FL and, in turn, indirectly affected their actual participation of FL activities. The connection between in-class and out-of-class eLearning activities also had a significant direct effect on student participation. These findings highlight the important role of instructional factors in promoting students' willingness to accept FL as a new learning approach in a traditional teacher-centered school subject.

Keywords: Chinese language learning, flipped learning, instructional factors, student acceptance

1. Introduction

Flipped learning (FL) is a type of blended learning that emphasizes the active role of learners by switching the traditional classroom instruction to pre-class home activities using educational technologies (Blau & Shamir-Inbal, 2017; Bond, 2020). A typical flipped classroom comprises two core components: out-of-class eLearning activities and face-to-face classroom learning (Bergmann & Sams, 2012; Rasheed et al., 2020; Strelan et al., 2020). When compared to the traditional classroom learning, FL has many benefits to student learning. The out-of-class eLearning component of FL allows students to decide the time, place, and frequency of learning based on their learning needs (Akçayır & Akçayır, 2018; Murillo-Zamorano et al., 2019; Rasheed et al., 2020). The in-class component of FL focuses on in-depth discussions and interactive activities, which helps students deepen their understanding of the learning content and enhance their higher-order thinking (Cheng et al., 2018; Lo et al., 2018; Sletten, 2017). Many studies have supported the effectiveness of FL in promoting students’ learning performance (Akçayır & Akçayır, 2018; Bond, 2020; Strelan et al., 2020), motivation (Kirmizi & Kömeç, 2019, Zou et al., 2022) and self-regulated learning (Blau & Shamir-Inbal, 2017; Çakiroğlu & Öztürk, 2021).

As the learning approach of eLearning is largely different from the traditional teacher-centered instruction, one major concern in technology enhanced learning research is students' acceptance and intention to use the technologies (Estriegana et al., 2019; Teo, 2019). As proposed by the technology acceptance model (TAM) (Davis, 1989; Venkatesh &
Davis, 2000), users’ perception of the usefulness and ease of use of a new technology significantly affects their intentional and actual use of the technology. In addition to perceived usefulness and ease of use, some researchers have proposed to add perceived enjoyment as another important indicator of student acceptance to represent students’ internal motivation to use the technology (Padilla-MeléNdez et al., 2013; Zacharis, 2012). Although TAM has been widely adopted to investigate user acceptance in many eLearning studies, only a few studies applied this model in understanding students’ acceptance of FL (Chen Hsieh et al., 2017; Haghighi et al., 2019; Padilla-MeléNdez et al., 2013). Most FL studies only explored students’ general attitudes towards FL, and inconsistent findings were found across different studies. While some studies indicated that students held positive attitudes toward FL (Akçayır & Akçayır, 2018; Kirmizi & Kömeç, 2019; Wang, 2016), some found students preferred traditional instruction to FL because of the extra workload and lacking teacher guidance in out-of-class eLearning (Akçayır & Akçayır, 2018; Bond, 2020; Lo & Hew, 2017).

Instructional design and implementation pf FL have been posited as significant determinants that affect students’ perception and acceptance of technologies and consequently affect their participation of FL activities (Dianati et al., 2022; Jiang et al., 2021). Regarding the out-of-class eLearning component of FL, many studies revealed that the quality of online teaching videos and other eLearning materials and activities would affect students’ willingness to participate in out-of-class eLearning (Akçayır & Akçayır, 2018; Bond, 2020; Haghighi et al., 2019; Lau & Keung, 2021; Rasheed et al., 2020). In addition, explicit instruction and constant feedbacks from teachers are important to support students’ independent learning in out-of-class eLearning activities, which can increase their acceptance and use of FL (Dianati et al., 2022; Zou et al., 2020). Researchers have also proposed some useful guidelines for the design and implementation of in-class learning activities of FL. First, there should be a close connection between out-of-class eLearning and in-class learning activities (Bond, 2020; Ng, 2018; Rasheed et al., 2020). Second, student-centered and cooperative learning activities should be designed to facilitate student interactions and cultivate students’ high-order cognitive abilities (Çakiroğlu & Öztürk, 2021; Haghighi et al., 2019; Lo & Hew, 2017; Strelan et al., 2020). Finally, teacher should also provide support to facilitate student involvement during the higher-order and interactive learning activities (Chen Hsieh et al., 2017; Sletten, 2017; Zou et al., 2020).

Compared with other subject areas, only a few FL studies were conducted in the area of language learning and were predominated by studies in the field of learning English as a second/foreign language (Turan & Akdag-Cimen, 2020; Zou, 2022). While findings of previous EFL/ESL studies supported the usefulness of FL in enhancing students’ language learning, certain difficulties and challenges were found when implementing FL in language classes. After reviewing 43 studies on using FL to promote English learning, Turan and Akdag-Cimen (2020) concluded that although FL could facilitate students’ interactions and overall language performance, it increased the workload and technical problems for both teachers and students. Studies on using FL to enhance language learning usually only measured students’ general perceptions or attitudes of FL (e.g., Dianati et al, 2022; Kirmizi & Kömeç, 2019). Only a few studies adopted TAM as a comprehensive framework to examine students’ acceptance of FL. For example, the results of TAM questionnaire in the study of Chen Hsieh et al (2017) indicated that students’ perceived usefulness, ease of use, and system characteristics were significant predictors of their intention of using FL to learn English. Chen Hsieh et al’s (2017) study was also the only study that included instructional factors in understanding students’ acceptance of FL in language learning. However, system characteristics only represented one component of FL design. Up to the present, no exiting studies have comprehensively examined the effects of different instructional factors on students’ acceptance in the subject area of language learning.

2. The Present Study
The study aimed to investigate the relations between instructional factors and students’ acceptance and use of FL in the context of L1 Chinese language learning. As shown in the above literature review, while FL has numerous positive effects on students’ learning, there are significant challenges in students’ acceptance and use of this new learning approach. To increase students’ acceptance and use of FL, the quality of the instructional design and implementation of FL is critically important. Moreover, most FL studies on language learning were conducted in the context of ESL/EFL. Influenced by Chinese Confucian culture, Chinese language teachers are used to adopt the traditional teacher-centered approach (Lau, 2013) and not familiar with eLearning (Turan & Akdag-Cimen, 2019). Thus, whether the design and implementation of FL in Chinese language classes adheres to the instructional principles of FL and how these instructional factors may affect students’ acceptance and use of FL in this subject area need to be further explored.

Specifically, the study sought answers for the following three research questions:

RQ1: How did students perceive the instructional design and implementation of FL in their Chinese language classes?
RQ2: What were the level of students’ acceptance and use of FL in Chinese language learning?
RQ3: Were students’ acceptance and use of FL significantly related to the instructional design and implementation of FL in their Chinese language classes?

3. Methodology

3.1 Participants

Data of this study was drawn from a large-scale survey study entitled “Hong Kong students’ acceptance of using FL in Chinese language subject and influencing factors” led by the first author. To ensure the representativeness of the sample, a similar number of students with different achievement levels were invited to participate in the study on a voluntary basis. The sample of this study consisted of 2160 students (1085 boys, 1017 girls, and 58 of unreported gender) between the ages of 12 and 17 years (mean = 14.55 years, SD = 1.62) from ten secondary schools in Hong Kong. Among these students, 724 were come from schools with high-achieving students, 742 from schools with moderate-achieving students, and 694 from schools with low-achieving students.

3.2 Measures

The study adopted two questionnaires to measure students’ perception of the instructional design and implementation of FL in their Chinese language classes and their acceptance and use of FL. Reliability estimates of the two questionnaires are shown in Table 1.

3.2.1 Instructional Design and Implementation of Flipped Learning Questionnaire (IFLQ)

This questionnaire was designed based on the instructional principles of FL and studies on the effects of instructional factors on the effectiveness of FL (Akçayır & Akçayır, 2018; Bergmann & Sams, 2012; Bond, 2020; Çakiroğlu & Öztürk, 2021; Lo & Hew, 2017; Lo et al., 2018; Ng, 2018; Rasheed et al., 2020; Zou et al., 2020). It consisted of 36 items with five subscales. Two of the subscales focused on the out-of-class eLearning component of FL measuring students’ perception of the quality of the eLearning materials and activities (OL_Q) and teacher support (OL_T) they received for out-of-class eLearning. Three of the subscales focused on the in-class learning component of FL measuring students’ perception of the connection between out-of-class eLearning and in-class learning (CL_C), the quality of the in-class learning materials and activities (CL_Q) and teacher support (CL_T). Students were asked to rate how similar was the description of each item with the instructional design and implementation of FL in their Chinese language classes on a 5-point Likert scale.
3.2.2 Technology Acceptance Model Questionnaire (TAMQ)

This questionnaire was adapted from the validated TAM questionnaires used in previous studies (Davis, 1989; Estriegana et al., 2019; Padilla-MeléNdez et al., 2013; Scherer et al., 2019; Teo, 2019; Venkatesh & Davis, 2000). The wordings of some items were revised to fit the context of Chinese language learning. The questionnaire consisted of 14 items with four subscales. Besides the two core components of TAM, perceived usefulness (PE) and perceived ease of use (PEU), perceived enjoyment (PEN) was added because of its importance in determining students’ acceptance of new technologies in the school setting (Lau & Keung, 2021; Padilla-MeléNdez et al., 2013). Students’ frequency of participating in the FL activities in their Chinese language classes was used to replace the “intention of use” variable in TAM to reflect their actual use (AU) (Estriegana et al., 2019; Zacharis, 2012). Students were asked to rate the level of their agreement with each item on a 5-point Likert scale.

Table 1. Descriptive statistics and reliability estimates for the two questionnaires

<table>
<thead>
<tr>
<th>Questionnaire/Subscale</th>
<th>No. of Items</th>
<th>Cronbach’s α</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructional Design and Implementation of Flipped Learning Questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out-of-class eLearning: Quality of Learning Materials &amp; Activities (OL_Q)</td>
<td>12</td>
<td>.95</td>
<td>3.63</td>
<td>.75</td>
</tr>
<tr>
<td>Out-of-class eLearning: Teacher Support (OL_T)</td>
<td>6</td>
<td>.92</td>
<td>3.65</td>
<td>.82</td>
</tr>
<tr>
<td>In-class Learning: Quality of Learning Materials &amp; Activities (CL_Q)</td>
<td>8</td>
<td>.95</td>
<td>3.72</td>
<td>.79</td>
</tr>
<tr>
<td>In-class Learning: Teacher Support (CL_T)</td>
<td>5</td>
<td>.92</td>
<td>3.68</td>
<td>.81</td>
</tr>
<tr>
<td>Connection between Out-of-class eLearning &amp; In-class Learning (CL_C)</td>
<td>5</td>
<td>.92</td>
<td>3.68</td>
<td>.81</td>
</tr>
<tr>
<td>Technology Acceptance Model Questionnaire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Effectiveness (PE)</td>
<td>5</td>
<td>.94</td>
<td>3.58</td>
<td>.84</td>
</tr>
<tr>
<td>Perceived Ease of Use (PEU)</td>
<td>3</td>
<td>.87</td>
<td>3.65</td>
<td>.85</td>
</tr>
<tr>
<td>Perceived Enjoyment (PEN)</td>
<td>3</td>
<td>.92</td>
<td>3.47</td>
<td>.97</td>
</tr>
<tr>
<td>Actual Use (AU)</td>
<td>3</td>
<td>.86</td>
<td>3.25</td>
<td>.84</td>
</tr>
</tbody>
</table>

3.3 Procedures and Data Analysis Plan

The initial version of the two questionnaires was reviewed by two researchers and four experienced Chinese language teachers with good experiences on using FL. After making revisions based on their comments, the questionnaires were piloted in a secondary school with moderate-achieving students. Further refinements were made based on the results of reliability estimates and confirmatory factor analysis (CFA) of the pilot study.

The questionnaires were distributed to the participating schools in June to August 2022. Schools could choose to administer either the paper-based or online version of questionnaires to their students. All students were required to complete the questionnaires anonymously under the supervision of their class teachers in about 10 minutes. Standardized instructions for administration were prepared for teachers to ensure all schools followed the same procedures to administer the questionnaires.

Reliability estimates and CFA were conducted again in the main study to check the psychometric properties of the questionnaires. The findings indicated the reliability and factor structure of both questionnaires were good (see Table 1 and 2). Descriptive analysis was done to examine students’ perceptions of the instructional design and implementation of FL in their Chinese language classes (RQ1) and their acceptance and use of FL (RQ2). SEM was conducted to examine the relations between all latent variables of the study (RQ3).
Table 2. Goodness of fit for the measurement models and SEM model of the Study

<table>
<thead>
<tr>
<th>Measurement model</th>
<th>Goodness-of-fit index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>IFLQ (5-factor model)</td>
<td>3134.234***</td>
</tr>
<tr>
<td>TAMQ (4-factor model)</td>
<td>323.771***</td>
</tr>
<tr>
<td>SEM model</td>
<td>5232.987***</td>
</tr>
</tbody>
</table>

4. Results

4.1 Instructional Design and Implementation of FL in Chinese Language Classes

As shown in Table 1, the mean scores of all subscales of IFLQ were higher than the mid-point of the 5-point Likert scale. Among the five subscales, the mean score of CL_T was the highest, followed by CL_Q, CL_C, OL_T, and OL_Q. The findings indicated that from the perception of students, the design and implementation of FL in their Chinese language classes generally adhered to the instructional principles of FL. Students showed a slightly more positive perception on the in-class learning component of FL than the out-of-class eLearning component.

4.2 Student Acceptance and Use of FL in Chinese Language Learning

As shown in Table 1, the mean scores of all subscales of TAMQ were higher than the mid-point of the 5-point Likert scale. Among the three acceptance variables, the mean score of PEU was the highest, followed by PE and PEN. The findings indicated that students generally showed a positive attitude towards using FL in Chinese language learning. The mean score of AU was slightly higher than the mid-point of the 5-point Likert scale, indicating a moderate level of student participation in Chinese language FL activities.

4.3 Relations between Instructional Factors and Student Acceptance of FL in Chinese Language Learning

SEM was adopted to explore the complex relationships between different instructional factors and students’ acceptance and use of FL in Chinese language learning. Based on the literature reviewed above, the model hypothesized students' AU of FL as the dependent latent variable, the five subscales of IFLQ as independent latent variables, and the three subscales of TAMQ as mediators. Item scores were used as observed indicators for all latent variables in the model. The SEM findings indicated a good fit for the hypothesized model (see Table 2).

The results of SEM are shown in Figure 1. Regarding the relations between instructional factors and students’ acceptance of FL, CL_Q had the largest significant effects on PE, PEU, and PEN and OL_Q also had significant effects on these three acceptance variables. No significant relation was found between the other three instructional variables and student acceptance. Regarding the relation between students’ acceptance and use of FL, PU had the largest significant effect on AU and PEN also had a significant effect on AU. Among the five subscales of IFLQ, only CL_C had a significant direct effect on AU. CL_Q and OL_Q mainly exerted effects on AU indirectly through the three acceptance variables. The indirect effects of CL_Q and OL_Q on AU were .12 (p < .01) and .15 (p < .01), respectively.
This study investigated Hong Kong secondary school students’ acceptance and use of FL in the context of L1 Chinese language learning. Since FL is largely different from the traditional teacher-centered instruction, previous studies found that not all students adapted to this new learning approach (Akçayır & Akçayır, 2018; Blau & Shamir-Inbal, 2017; Bond, 2020). In contrast to the prediction of the study that this problem may be more obvious when FL is introduced to Chinese language classes which has long been dominated by the traditional teacher-centered and knowledge-based instruction (Lau, 2013), the findings indicated a moderate level of student participation in Chinese language FL activities and students generally showed a moderately high level of acceptance to this new learning approach.

Among the three acceptance variables, students had the most positive perception on ease of use. Since online learning has been adopted frequently in Hong Kong during the COVID-19 outbreak in the past three years, most Hong Kong secondary school teachers and students are familiar with various types of eLearning tools and platforms, which may help reduce students’ technical problems in FL. Students’ positive perceptions on the usefulness and enjoyment of FL are consistent with the positive effects of FL on students’ learning outcomes (Akçayır & Akçayır, 2018; Bond, 2020; Strelan et al., 2020) and motivation (Kirmizi & Kömeç, 2019; Zou et al., 2022) revealed in many previous studies. As posited by some language educators, FL move the teaching of a large amount of basic knowledge and lower-level language skills to better organized and self-paced pre-class self-learning and free class time for students to interact actively with their classmates to practice and apply their higher-order language skills, which consequently promote higher motivation and proficiency in language development (Li et al., 2022; Jiang et al., 2022; Turan & Akdag-Cimen, 2019). The study’s findings support that these proposed benefits of FL in language learning are also endorsed by Hong Kong students when FL is incorporated in their Chinese language classes.

Students’ positive perception on the design and implementation of FL in Chinese language classes refute the traditional view that Chinese language teachers tend to adopt teacher-centered instruction (Lau, 2013) and FL approach is seldom applied in Chinese
language teaching (Zou et al., 2022). This may be due to a greater emphasis of student-centered instruction in current Chinese language curriculum (HKCDC, 2017) and teachers’ more frequent use of eLearning tools and platform during the COVID-19 outbreak. Nevertheless, the study revealed that students held more positive perception on the in-class learning component of FL than the out-of-class eLearning component. It suggests that Hong Kong Chinese language teachers are still more capable in designing high quality learning materials and activities and supporting students’ learning during face-to-face classroom teaching than in the out-of-class eLearning context. Thus, more trainings should be provided for Chinese language teachers to enhance their professional skills in incorporating eLearning in Chinese language teaching.

In line with the previous TAM studies (Chen Hsieh et al., 2017; Estriegana et al., 2019; Haghighi et al., 2019; Padilla-MeléNdez et al., 2013; Zacharis, 2012), PE and PEN were found to have significant effects on AU. While the largest effect of PU on AU confirms that perceived usefulness is the most critical variable in TAM (Scherer et al., 2019), the significant relation between PEN and AU and the non-significant relation between PEU and AU suggest that young school students consider the feeling of enjoyment more than the difficulty level of the tasks when they determine whether they would like to participate in FL activities (Padilla-MeléNdez et al., 2013; Zacharis, 2012). The significant direct effects of the instructional factors on student acceptance of FL support the important role of instructional factors in enhancing students’ acceptance of using FL (Cheng, 2019; Dianati et al., 2022; Hsieh et al., 2017; Jiang et al., 2021). Among various instructional factors, CL_Q and OL_Q were found to have significant effects on all the three student acceptance variables and indirectly affected AU. These findings are consistent with the view that student-centered, higher-order, and interactive in-class activities, and interesting, diverse, and well-designed eLearning materials and activities can enhance students’ positive attitude towards FL and, in turn, increase their willingness to participate actively in FL activities (Akçayır & Akçayır, 2018; Blau & Shamir-Inbal, 2017; Bond, 2020; Haghighi et al., 2019; Lau & Keung, 2021; Rasheed et al., 2020). It is noteworthy that CL_C did not have any significant effect on student acceptance but it had a direct effect on AU. Since many students view out-of-class eLearning activities as extra work for them and are not willing to make pre-class preparation (Akçayır & Akçayır, 2018; Rashead et al., 2020), teachers’ effective use of follow-up activities is essential for encouraging students to complete pre-class learning tasks (Blau & Shamir-Inbal, 2017; Lau & Keung, 2021; Lo et al. 2018).

In conclusion, while previous FL studies have predominately focused on the effectiveness of FL on student learning, the study made a unique contribution to the FL research by verifying the relations between instructional factors, students’ acceptance, and their use of FL using more comprehensive measures of instructional factors and student acceptance. The study’s findings highlight the importance of designing high quality in-class and out-of-class eLearning materials and activates to increase students’ willingness to participate in FL activities. The study’s findings also confirm the mediating role of student acceptance between instructional factors and students’ actual participation in FC activities. Students’ positive perceptions on the design and implementation of FC in their Chinese language classes and their high level of acceptance of this new learning approach support the potential benefits of adopting FL to enhance student learning in this traditional teacher-dominated school subject. Lastly, it should be noted that the measure of instructional factors in the study were based only on students’ perception. The significant relations between instructional factors and student acceptance of FL should be replicated in future studies using direct measures of instructional design and implementation.

Acknowledgements

This study was supported by the Standing Committee on Language Education and Research (EDB(LE)/P&R/EL/203/2).
References


Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. DC: International Society for Technology in Education.


The AI-Supported Instructional Design in PBL Integrating Chinese Language Learning and Multimedia Creation

Satoko SUGIE a*

a Faculty of Humanities, Sapporo International University, Japan
* satoko-sugie@ts.siu.ac.jp

Abstract: The rapid development of AI (Artificial Intelligence) is enhancing the potential to accelerate a paradigm shift in foreign language teaching and learning. Instructional design must shift beyond knowledge learning and skills training to include more significant opportunities for practical application in authentic contexts. To do so, it is necessary to investigate the pedagogical possibilities of generative AI through specific educational practices. Thus this study intends to explore the potential of using generative AI to improve the efficiency and effectiveness of PBL (Problem/Project Based Learning) which integrates foreign language, information media, and professional education. This study suggests a PBL course design integrating Chinese language, tourism, and multimedia creation learning outcomes at a Japanese university, as a preliminary phase to compare and examine the teaching and learning processes between traditional methods and proactive AI-enabled methods. Through qualitative exploration of integrating generative AI into traditional instructional design, this paper considers how to organize the elements of teaching and learning for a renewed approach with greater efficiency and quality of learning outcomes.

Keywords: generative AI, Chinese as a foreign language, multimedia creation, instructional design, Project-Based Learning

1. Introduction

Japanese universities commonly teach introductory foreign language courses through lecture-style classes, where in-class activities are mainly individual knowledge learning. Since the outbreak of the COVID-19 pandemic, the use of ICT (Information and Communication Technology) in education has become somewhat more normal. From a broader perspective, blended and hybrid ICT-based teaching/learning has been identified as a worthy challenge to assist teachers in proficiency of ICT skills (Beatty, 2019). However, even during the pandemic, MEXT (Japan’s Ministry of Education, Culture, Sports, Science, and Technology) emphasized traditional, in-person instruction (MEXT, 2021a), expressing a hope that by 2023, there would be a return to traditional teaching (MEXT, 2021b).

In contrast to government hopes, the evolution of generative AI has been a societal paradigm shift that educators noticed. Generative AI supports more efficient education of vocabulary, grammar, and writing in foreign languages. Frequent updates have further enhanced accuracy, rendering AI more cost-effective (Kondo et al., 2023; Niño, 2020; Tsai, 2020). However, most traditional teachers doubt the improvement through AI, and think communication practice cannot be achieved without repeating physical lessons on vocabulary, grammar, and pronunciation (Onishi, 2008). Teachers from a behaviorist or communicative background trust far more in the methods they applied during their own success as foreign language learners. Studies investigating how to use generative AI are vital because the problem of ‘the Chasm’ (Moore, 1991) is equally true in the education field, and there may be a deep divide between early adopters and the majority of teachers on how best to apply such new technology.
2. Research Question

This study seeks to use the Japanese government’s tentative guidelines for the use of generative AI (MEXT, 2023). The seven points of that guideline can be summarized as follows.

| Table 1. The seven points of the tentative guidelines for the use of generative AI |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| (1) Encourage moral education   | use AI-generated information as a teaching tool for learners to discover its characteristics and limitations. |
| (2) Allow discussions on        | use Generative AI as a material for learners to think and discuss more actively.                                                   |
| Generative AI                   |                                                                                                                                  |
| (3) Support clear expression     | discover new perspectives on the results of summaries and discussions in group work with a generative AI.                         |
| and generalization of           |                                                                                                                                  |
| individual learners’ knowledge, |                                                                                                                                  |
| thoughts, and lived experiences  |                                                                                                                                  |
| (4) Optimize language assistance | use AI as a dialogue companion or improved dictionary to learn more natural and native-like expressions, searching for words and example sentences. |
| in practice exercises           |                                                                                                                                  |
| (5) Support better editing      | repeatedly use AI to revise sentences learners have written themselves to improve the quality.                                    |
| (6) Encourage advanced learning | use sophisticated programming through generative AI.                                                                               |
| (7) Support assessment          | test tools to assess problem-finding and problem-solving abilities using generative AI.                                           |

Since this study focuses on foreign language education, there are possibilities for learners to use guidelines 3, 4, and 5, and for teachers to use 7.

This study is action research to explore situational questions with a qualitative approach. Based on the instructional design of a class conducted by the author in 2023, learning activities and tools have been organized which could potentially apply generative AI. To examine the effectiveness of teaching and learning using generative AI compared to traditional teaching and learning strategies, it is necessary to identify what teaching and learning processes are used in traditional instructional design, starting with an exploration of the potential for improvement. The research question of this study is as follows: How could AI enhance a PBL course integrating foreign language learning and multimedia creation?

3. Methods

In this study, new viewpoints on how to integrate generative AI were examined in comparison to traditional instructional design, through subjective observations based on this author’s years of teaching experience. Analysis was conducted on the design of a spring semester course (15 weeks), considering the strategies, outcomes and likely challenges in lesson design and learning tasks.

3.1 Class Setting

The learners were twenty-three first-year Japanese university students who have studied basic Chinese as a foreign Language for less than one year. The class is composed of a mixture of Humanities and Science majors. The learning environment is a face-to-face class in a general classroom, where students therefore bring their own devices such as laptops and smartphones. The network environment is connected to the university's Wi-Fi. Learners use the LMS (Learning Management System) for class communication, browsing learning materials, submission of assignments, and reviewing teacher feedback. Grading is based on the submission of in-class and out-of-class assignments, in-class discussions, the submission of videos on learning outcomes, and a final presentation in Chinese.
3.2 Learning Goals and Activities

The class has two main themes. The first is cross-cultural understanding and vocabulary learning necessary for welcoming Chinese FIT (Foreign Independent Tour) visitors to Hokkaido, Japan’s Northernmost province. For this purpose, students learn words, phrases, and conversational expressions used for tourism hospitality through a set textbook. Moreover interactive learning happens as international students answer learners’ questions about Chinese culture. Second, the students create a tourism promotional video in Chinese. For this purpose, students make a video in groups that discovers, co-creates, and promotes local attractions from the learner’s perspective. This requires extracurricular collaboration. The three learning goals are as follows: (1) To be able to address and answer simple Chinese questions in the tourism industry and hospitality situations, (2) Understand the norms, beliefs, and culture of Chinese visitors to Japan from a cross-cultural perspective, and (3) To be able to utilize Chinese through tools including distance learning, automatic voice translators, and video editing applications.

4. Data analysis

A qualitative approach was adopted involving analysis of lesson plans, teaching experience and evaluation of possible student outcomes. The comparison of traditional and AI-enhanced course designs, including learner and teacher outcomes are shown in Figure 1. The conventional work of teachers and learners across the usual fifteen-week course was summarized, and then potential AI-supported alternatives were devised.

<table>
<thead>
<tr>
<th>Learning activities</th>
<th>Learners’ work</th>
<th>Teacher’s work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sess.2: Sharing of Chinese name lists and roll call (in every class). Learning and practicing Chinese pronunciation and pinyin.</strong></td>
<td><strong>Independent work</strong>&lt;br&gt;- Listen to their own name in Chinese and respond in Chinese.&lt;br&gt;- Read aloud the words and sentences instructed by the teacher.&lt;br&gt;- Pair/group work&lt;br&gt;- Practice customer and clerk conversation.</td>
<td><strong>[PRON] Input words and sentences in the textbook via AI chat and translation Applications and practice the pronunciation of the synthetic voice.</strong></td>
</tr>
<tr>
<td><strong>Sess.3: Sharing the questions and answers about China and Chinese culture. Writing a short essay about self-introduction.</strong></td>
<td><strong>Independent work</strong>&lt;br&gt;- Summarize international students’ answers to questions from Japanese students.&lt;br&gt;- Write a detailed self-introduction in Japanese.&lt;br&gt;- Translate the Japanese self-introduction into Chinese.&lt;br&gt;- Pair/group work&lt;br&gt;- Speech practice of self-introduction.</td>
<td><strong>[INF-S] Summarize international students’ answers.</strong>&lt;br&gt;- [LI/TR] Translate Japanese into Chinese. <strong>[PRON] Check and practice correct pronunciation using the audio function of the dictionary and translation Applications.</strong> <strong>[DE/FAQ] Practice dialogue and FAQ sessions with AI.</strong></td>
</tr>
<tr>
<td><strong>Theme 1: cross-cultural understanding and vocabulary learning necessary for welcoming Chinese FIT visitors</strong></td>
<td></td>
<td><strong>[LI/TR] Case-based discussion:</strong> Students discuss and answer questions in Chinese. <strong>[PRON] Correct student writing and provide feedback with revisions and comments on improvements.</strong> <strong>Listen to the student’s speech and advise revisions.</strong>&lt;br&gt;- <strong>[PRON] Let learners waiting their turn for tutoring practice pronunciation, writing, and translation.</strong> <strong>[PRON] Let learners waiting their turn for tutoring practice pronunciation using a voice-activated pronunciation practice system.</strong></td>
</tr>
<tr>
<td><strong>Teacher’s work</strong>&lt;br&gt;- Read and correct Chinese self-introductions. <strong>[PRON] Listen to a Chinese self-introduction and correct pronunciation.</strong> <strong>Practice accurate pronunciation through group and individual instruction.</strong> <strong>[LI/TR] Consider more appropriate expressions with AI translation.</strong></td>
<td><strong>AI technology will be able to automatically analyze the pronunciation characteristics of individual learners and provide practice exercises and corrective tips.</strong></td>
<td></td>
</tr>
<tr>
<td>Sess.4:</td>
<td>Sharing questions and answers about China and Chinese culture. Learning how to translate foreign words into Chinese. Chinese pronunciation practice.</td>
<td>Independent work</td>
</tr>
<tr>
<td>Sess.5-8:</td>
<td>Sharing FAQ about China and Chinese culture. Learning and comparing the Basic Characteristics of the Chinese people.</td>
<td>Independent work</td>
</tr>
<tr>
<td>Theme 2: the students will create a tourism promotion video using the Chinese language</td>
<td>Sess.9:</td>
<td>Writing a proposal for a Hokkaido short trip.</td>
</tr>
<tr>
<td>Sess.10:</td>
<td>Creating a video composition chart.</td>
<td>pair/group work</td>
</tr>
<tr>
<td>Sess.11: Video shooting planning. Learning how to use video editing APPs.</td>
<td>Video editing and feedback on comments.</td>
<td>pair/group work</td>
</tr>
<tr>
<td>Sess.12: Video shooting planning. Learning how to use video editing APPs.</td>
<td>Video editing and feedback on comments.</td>
<td>pair/group work</td>
</tr>
<tr>
<td>Sess.13: Translation of Chinese ticker and narration dialogue. Practicing introduction of a video overview.</td>
<td></td>
<td>pair/group work</td>
</tr>
<tr>
<td>Sess.14: Revising the video. Submit video assignments.</td>
<td></td>
<td>pair/group work</td>
</tr>
<tr>
<td>Sess.15: Presentation and mutual evaluation.</td>
<td></td>
<td>pair/group work</td>
</tr>
</tbody>
</table>

Figure 1. Comparison of AI-supported instructional design, activities, and tasks
5. Discussion

The improvements over the traditional teaching and learning process as outlined in Figure 1 can be categorized into the following seven aims. The acronyms shown in [ ] in AI-supported elements in Fig.1 represent the same seven elements.

1. [LI/TR] Language input and translation: Japanese to Chinese/Chinese to Japanese translation by text, Chinese input by voice
2. [PRON] Pronunciation Practice: Input (listen for accurate pronunciation) and output (correct pronunciation)
4. [INF-CL] Information collection
5. [INF-CT] Information categorization
6. [INF-S] Information summary

The following are examples of tools and applications that could be used for teaching and learning activities. Most are free, but if there is a fee, it would be necessary to work with the organization’s instructional management department or for individual teachers to manage the expense of using student licenses and setting up registration management.

1. **Translation**: DeepL, Google Translate, ChatGPT, Bing and Bard
2. **Pronunciation practice**: Ondoku-san (ondoku3.com/ja), Microsoft Azure (speech.microsoft.com/audiocontentcreation), Ondoku Chinese (ondoku.org)
3. **Oral presentation practice**: RESEMBLE.AI (English to Chinese conversion) (www.resemble.ai)
5. **Image/video creation**: Microsoft Designer (designer.microsoft.com), Google Imagen Video (imagen.research.google/video), Meta Make a Video (makeavideo.studio), Gen-2 (research.runwayml.com/gen2), Stable Diffusion (stablediffusionweb.com), Scribble Diffusion (scribblediffusion.com)

We have recommended the use of such tools for (1) machine translation and (2) speech recognition for text input and pronunciation correction among traditional learning activities. However, they were only used during extracurricular individual learning. To increase the ratio of use in cooperative and autonomous learning within the classroom and realize more efficient activities, it is necessary to consider the optimal teaching methods which allow learners to know which tools to use in which situations, and how to practice their skills.

The following are points to be considered in AI-supported instructional design. Generative AI is still in the early trial-and-error stage of development so teachers must determine which tools are optimal for which tasks. Teachers need to continue to improve the task design and instruction of the most appropriate AI-enabled pedagogical methods through implementation and review of practical data. The cost of learners’ low proficiency in tool operation itself will squeeze time for cooperative learning activities and interactive feedback in the classroom. It is thus necessary to improve AI literacy in the form of cross-curricular activities using on-demand videos and flipped learning. Likewise, it is preferable to establish a team-based instructional design and class management system.

6. Conclusion and Future Challenges

This study examined the potential of using generative AI in PBL, which combines foreign language, ICT utilization, and professional education. Seven key areas were identified where generative AI could improve teaching and learning efficiency. However, human teachers are
still needed to provide integrated instruction based on their experience in dealing with each
diverse learner. In conversations and collaborative work, teachers must provide pedagogical
instruction that fosters autonomous, interactive, and in-depth learning. As well as blended or
hybrid ICT-based teaching/learning, the AI-supported instructional design also requires
establishing a cross-subject educational system by networking with information course
teachers and other advanced educational practitioners. It is essential to reconfirm that
teachers are core actors for instructional design, and then recognize how best to allocate
teaching time to AI-enabled resources. In addition to that, I believe that it is necessary to
maximize the allocation of learning support that only a human teacher can provide.

To test the implications of this study, future studies must add quantitative practical
data on implementation of the suggested instructional design. As this field grows, it is vital to
see more sustained comparison of the learning process and outcomes between traditional
learning methods and activities fully utilizing a generative AI.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H01276.

References

Hybrid-Flexible Course Design: Implementing Student-Directed Hybrid Classes, EdTech Books.
https://edtechbooks.org/hyflex/

Katelay, A., Ide, K., & Kishimoto, M. (2023, April 13). Seisei AI (Generative AI) no ronri-teki hōteki
shakai-teki kadai (ELSI) ronten no gaikan [Overview of ethical, legal, and social issues (ELSI) in

Kondo, Y., Kimura, S., Sakaba, H., Toyoshima, C., Nakanan, M., Yamashita, M., Yamanaka, T.
(2023). AI kikai honyaku no eigo seika jugyou he no daikibo donyu to sono kadai [Feasibility Study
on Large-scale Introduction of Neural Machine Translation into PBL English Classes],
Proceedings of the CIEC Spring Conference2023, 14, 41-44.

University Press.

Ministry of Education Culture, Sports, Science and Technology [MEXT]. (2021a). Reiwa 3 nendo koki
no daigaku tou ni okeru jugyo no jisshi hoshin tou ni tuite [Notes on the Conduct of Classes at
Universities and Other Institutions in Academic Year 2021 and Countermeasures against New
Coronavirus Infections, etc. (Notification)]. https://www.mext.go.jp/content/20210305-
mxt_kouhou01-000004520-02.pdf

Conduct of Classes at Universities and Other Institutions in the First Half of the 2021 Academic
Year. https://www.mext.go.jp/content/20210702-mxt_kouhou01-000004520_2.pdf

for the Utilization of Generative AI in the Primary and Secondary Education Levels.
https://www.mext.go.jp/a_menu/other/mext_02412.html

Customers, HarperCollins.


Kaihatsu no kokoromi [Direction of Second Language Education from Now Onward: Efforts in
Developing a Unified Textbook for Chinese]. Kinki University Department of Language Education

Sugie, S. (2023, March 18). Kanko Chugokugo Maruchimedia wo tougou shita purojekuto gata
gakushu no seika to kadai [Achievements and Challenges of Project-Based Learning Integrating
Tourism, Chinese, and Multimedia Education], 20th Association for e-Learning Language
Education Annual Conference 2023, Ritsumeikan University, Umeda campus.
https://researchmap.jp/ssugie/presentations/41770097

Tsai, S. C. (2020). Chinese students’ perceptions of using Google Translate as a translingual CALL
tool in EFL writing, Computer Assisted Language Learning, 35(5-6)1250-1272.
https://doi.org/10.1080/09588221.2020.1799412
Experimental Comparison of Promotion Effect for EFL Reading Comprehension between Conventional Summarization and Toulmin Argument Reconstruction

Banni Satria ANDOKOa*, Rosa Andrie ASMARAb, Vivin Ayu LESTARIa, Deasy Sandhya Elya IKAWATIa, Ariief PRASETYOa, Tsukasa HIRASHIMAb & Yusuke HAYASHIb

a Information Technology Dept, State Polytechnic of Malang, Indonesia
b Graduate School of Advanced Science and Engineering, Hiroshima University, Japan
*ando@polinema.ac.id

Abstract: Summarization is a conventional task to promote reading comprehension. However, because it is impossible to diagnose the summary immediately and accurately, it is not easy to realize formative feedback for ongoing reading comprehension. To solve this issue, we have proposed Viat-map reconstruction as a task for promotion of reading comprehension. Viat-map is a three-component (Data, Warrant and Claim) representation of the Toulmin Argument. Reconstruction means that a learner is required to reconstruct a Viat-map prepared by an expert by using provided components. By comparing the reconstructed map with the original map, it is possible to realize immediate and accurate diagnosis and ongoing formative feedback for the reading comprehension. We have conducted a comparative experiment of promotion effect for EFL reading comprehension between conventional summarization and Viat-map reconstruction. Fifty-eight second year students of Information Technology Department from State Polytechnic of Malang, Indonesia were involved in the experiments. All of them are using English as Foreign Language. The experiment was conducted in an English course subject for three weeks. Experiment and control group were created equally based on the initial test of reading comprehension test. The experiment group was using Viat-map, while control group is using conventional summarization method. The result of the experiment was analyzed by using ANOVA statistical analysis and Holm’s Sequentially Rejective Bonferroni Procedure for POST Hoc analysis. The result shows that Viat-map reconstruction overachieved summarization and both groups can maintain the memory one week after the lesson. However, the delay-post test score for experiment group is higher than control group.

Keywords: English Reading Comprehension, Toulmin Argument, Viat-map, EFL

1. Introduction

English has become one of the most popular languages in the world and is recognized as the most widespread language of communication in the fields of education, technology, business, diplomacy, science, and sports, as well as the service-oriented high-tech industries and many other fields (Zhang, 2011). For every student, regardless of their English proficiency, it is important that they can implement their writing, reading, speaking, and listening skills in their academic or scientific environment as well as their social life. (Dehham et al., 2022). In formal education, studying English as English as Foreign Language (EFL) students, took more account on reading abilities, normally because it is the main point of contact with the target language. English knowledge as international language of communication, is becoming more necessary for many adult university learners (Allard & Mizoguchi, 2021).

Reading is one of the important activities to gain knowledge and information. In other words, reading is one of the gateways to gaining an understanding. Because reading is an active and continuous process that involves people with reading material to build meaning or information (Carrell, 1989). Reading is also interpreted as an active task in which the reader
makes choices from specific words derived from the text and relates them to the situational context to build a model of meaning that describes the meaning that is the same as the meaning intended by the writer (Dole et al., 1991).

Sentence structures in reading materials are generally presented to readers in the form of linear sentences without any good directions to assist the reader in recognizing the logical structure of the text (Eftekhari et al., 2016; Eftekhari & Sotoudehnama, 2018). The condition creates a huge confusion, especially for EFL students. Logical structure of text could assist students to promote an autonomous learning and enhance the depth of learning (Andoko et al., 2020). Toulmin Argument could also be applied to practical teaching (Yang, 2022b, 2022a).

The motivation of this study is assisting students to understand the reading material by using Toulmin Arguments strategy and Viat-map application. Research questions of this study are:

1. Do students who use Viat-map application have a better understanding compared to ones who use conventional summarization method?
2. Do students who use Viat-map application maintain their understanding compared to ones who use conventional summarization method?

Several procedures will be conducted to answer the research question, measuring the level of students understanding and memory retention when using the Viat-map application compared to the ones who are using the conventional summarization method. First, we will compare the pre-test score with the post-test score from each group to find any improvement and then comparing the post-test score between groups to find any differences as the representative of students’ understanding. Secondly, we will compare the delayed test score with the post-test score from each group to find any changes and then compare the delayed test between groups as the representative of students’ memory retention.

2. Literature Review and Related Research

2.1 Toulmin Argument

Toulmin Arguments is a form of implementation of a graphical strategy based on the Argumentation model. The argumentation model can give confidence and influence on the reader because it presents logical and strong reasons to prove the truth of an opinion based on data and facts (Hitchcock, 2005; Junaidah Januin, 2021). Toulmin's argument is a complete argumentation pattern and has a good understanding concept, Toulmin's argument has a description of the learning process which consists of claims, qualifiers, ground, warrants, rebuttals, and backing (Magalhães, 2020; Toulmin, 2003). Toulmin's argument is formed by 3 basic elements, namely Claim, Ground and Warrant (Figure 1). Claim is a sentence, information, conclusion to be conveyed. Ground are facts, evidence or data that become a reference for a Claim, while Warrant is a sentence that connects Ground with a Claim. These three must be formed into a logical framework to provide a strong understanding of literacy (Andoko et al., 2022).

![Figure 1. Toulmin Argument concept.](image-url)
2.2 Viat-map

Viat-map is an application developed to provide a deep understanding of literacy. Using it for Reading Comprehension in English for EFL students shows a promising result (Andoko et al., 2022; Rismanto et al., 2021). The main concept is to form a logical framework of text by adopting Toulmin's argument. It intends to emphasize reasoning so students can construct a strong understanding of information within the text. There are three main phases of Viat-map (Figure 2):

1. Teacher’s logical map: in this phase, teacher is responsible to create several exercises by selecting important sentences as claim and provide one correct evidence or fact as ground along with two incorrect evidences. Teacher also provides one correct relation between sentences and evidence (warrant) followed by two incorrect relations.

2. Students’ working space: in this phase, students are constructing provided exercise by the teacher by selecting the correct ground and warrant, they cannot move to the next claim until they find the correct ground and warrant.

3. Teacher’s overlapping analysis feedback: in this phase, teacher can provide feedback by using the overlapping analysis. Teacher can emphasize more feedback by following the number present in each line of answer. The number consists of two different pieces of information, the first number stands for the number of mistakes made by the students, the second number is for the total number of students who made the mistakes. For further information, the teacher can select the line and it will show the information of the students and the number of mistakes made by each student.

![Figure 2. The phases of Viat-map application.](image-url)
4. Methods

4.1 Experimental Setting

The study took place in October 2022 in an English subject Class of second year Information Engineering Department student of State Polytechnic of Malang. All the students were using English as Foreign Language. There were fifty-eight students involved coming from two different classes. The student never used and familiar with Toulmin argument before. The study was conducted in three weeks where the first week is used to introduce them with Toulmin Argument concept, Viat-map application and to conduct the initial test used to categorize students into a control or experiment group. Each class was divided into two groups to ensure the consistency of the experiment. The next following week was used to conduct the experiment. The third week was used for delayed test. During the experiment, the teacher never give explanation or feedback. Her task was providing the reading material, creating the Toulmin Argument and evaluating the score.

Control groups would use the conventional summarization method where they were creating a summary from the reading material to simplify or to emphasize the most important sentence and words. While control group would be using the Viat-map application to construct their understanding of the reading material. The initial test was using three different reading comprehension material without any treatment to ensure that we get their basic score of reading comprehension test. The experimental design used in the second week was carefully conducted to ensure that we had a controlled environment to avoid unbalanced treatment. The second reading material used of the experiment was using the underlined important sentences so that the control group receive the same hint as the experimental group. The design for the experiment is shown in Figure 3.

![Figure 3. Experimental design.](image)

4.2 Measurement and Analysis

The scoring measurement used in this experiment was using a normal calculation score, each correct answer will receive 1 point. The test material for each group for pre, post and delay-test are using the same material. Students never know their mistakes since the test result is never exposed to them until the experiment is over. The test material consists of 10 multiple choices. Each correct answer will be given 1 point and 0 point for incorrect. Total score is the summary of correct answer. The Analysis was using ANOVA (Analysis of Variance). Considering the students are having the same background, age, and the percentage of gender
is almost the same where Male is 54% while female is 46%. Post hoc analysis using Holm’s Sequentially Rejective Bonferroni Procedure was also conducted to explain the ANOVA result in detail.

5. Result

This section will analyze the reading comprehension score from each group. The first step of analysis is the mean analysis for each group for every test (Table 1). The result shows that the standard deviation is smaller than mean. It indicates that each data is clustered around each mean. The second step is checking the Sphericity Test and Epsilons before conducting the ANOVA calculation (Table 2). Based on the result, the sphericity is violated (p-value = 0.0001; p < .001), it is occurred because the number of data sample is small. Greenhouse-Geisser (GG), Huynh-Feldt-Lecoutre (HF) and Chi-Muller (CM) correction are used to correct the violation of the assumption of sphericity, it shows that the variance of differences between groups are equal (GG = 0.7816; HF = 0.8179; CM = 0.8148) and the ANOVA calculation can be conducted.

Table 1. Mean Analysis for each group and test

<table>
<thead>
<tr>
<th>GROUP</th>
<th>test</th>
<th>n</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>El.InitTest</td>
<td>29</td>
<td>5.6897</td>
<td>1.6280</td>
</tr>
<tr>
<td>Control</td>
<td>El.PreTest</td>
<td>29</td>
<td>6.1034</td>
<td>1.8774</td>
</tr>
<tr>
<td>Control</td>
<td>El.PostTest</td>
<td>29</td>
<td>5.8621</td>
<td>1.7672</td>
</tr>
<tr>
<td>Control</td>
<td>El.DelayedTest</td>
<td>29</td>
<td>5.7931</td>
<td>2.4550</td>
</tr>
<tr>
<td>Experiment</td>
<td>El.InitTest</td>
<td>29</td>
<td>5.5517</td>
<td>1.8242</td>
</tr>
<tr>
<td>Experiment</td>
<td>El.PreTest</td>
<td>29</td>
<td>5.9655</td>
<td>1.4264</td>
</tr>
<tr>
<td>Experiment</td>
<td>El.PostTest</td>
<td>29</td>
<td>6.9655</td>
<td>1.5920</td>
</tr>
<tr>
<td>Experiment</td>
<td>El.DelayedTest</td>
<td>29</td>
<td>6.8966</td>
<td>1.7596</td>
</tr>
</tbody>
</table>

Table 2. Mendoza’s Multisample Sphericity Test and Epsilons

<table>
<thead>
<tr>
<th>Effect</th>
<th>Lambda</th>
<th>Approx. Chi df</th>
<th>p-value</th>
<th>LB</th>
<th>GG</th>
<th>HF</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>0.0000</td>
<td>37.2606 11</td>
<td>0.0001 ***</td>
<td>0.3333</td>
<td>0.7816</td>
<td>0.8179</td>
<td>0.8148</td>
</tr>
</tbody>
</table>

LB = lower.bound, GG = Greenhouse-Geisser HF = Huynh-Feldt-Lecoutre, CM = Chi-Muller

After checking the sphericity, ANOVA calculation was conducted, and the result can be seen in Table 3. The result indicates that there is no significant difference found between groups for all combined test scores (p-value = 0.2271). However, there are significant different found for test within the groups (p-value = 0.0012; p < .01) and each test between groups (p-value = 0.0013; p < .01). Based on the ANOVA calculation results, there are significant differences found for the test and each group test. However, the detailed information cannot be determined and needs to conduct a post hoc analysis.

Table 3. ANOVA calculation result for each group and tests.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>13.5172</td>
<td>1</td>
<td>13.5172</td>
<td>1.4916</td>
<td>0.2271 ns</td>
</tr>
<tr>
<td>S X GROUP</td>
<td>507.5000</td>
<td>56</td>
<td>9.0625</td>
<td></td>
<td></td>
</tr>
<tr>
<td>test</td>
<td>22.7586</td>
<td>3</td>
<td>7.5862</td>
<td>5.5558</td>
<td>0.0012 **</td>
</tr>
<tr>
<td>GROUP X test</td>
<td>22.3448</td>
<td>3</td>
<td>7.4483</td>
<td>5.4548</td>
<td>0.0013 **</td>
</tr>
<tr>
<td>S X GROUP X test</td>
<td>229.3966</td>
<td>168</td>
<td>1.3655</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Holm's Sequentially Rejective Bonferroni Procedure analysis for "tests" and "GROUP X test" need to be conducted to show a better result. The alpha level used is 0.05. Firstly, mean analysis needs to be conducted to make sure that the data is clustered around the mean (Table 4). The result shows that all the standard deviations are below the mean of each test. Multiple Comparison for "test" then conducted (Table 5). It shows that the combination of post-test score is higher than the combination of initial test while other remain equal.

Table 4. *Mean Analysis for test in groups.*

<table>
<thead>
<tr>
<th>test</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>El.InitTest</td>
<td>58</td>
<td>5.6207</td>
<td>1.7150</td>
</tr>
<tr>
<td>El.PreTest</td>
<td>58</td>
<td>6.0345</td>
<td>1.6540</td>
</tr>
<tr>
<td>El.PostTest</td>
<td>58</td>
<td>6.4138</td>
<td>1.7575</td>
</tr>
<tr>
<td>El.DelayedTest</td>
<td>58</td>
<td>6.3448</td>
<td>2.1889</td>
</tr>
</tbody>
</table>

Table 5. *MULTIPLE COMPARISON for "test"*

<table>
<thead>
<tr>
<th>Pair</th>
<th>Diff</th>
<th>t-value</th>
<th>Df</th>
<th>p</th>
<th>Adj.p</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1.InitTest-E1.PostTest</td>
<td>-0.7931</td>
<td>3.5603</td>
<td>56</td>
<td>0.0008</td>
<td>0.0046</td>
</tr>
<tr>
<td>E1.InitTest-E1.DelayedTest</td>
<td>-0.7241</td>
<td>2.6037</td>
<td>56</td>
<td>0.0118</td>
<td>0.0589</td>
</tr>
<tr>
<td>E1.PreTest-E1.PostTest</td>
<td>-0.3793</td>
<td>2.3405</td>
<td>56</td>
<td>0.0228</td>
<td>0.0914</td>
</tr>
<tr>
<td>E1.InitTest-E1.PreTest</td>
<td>-0.4138</td>
<td>2.0217</td>
<td>56</td>
<td>0.0480</td>
<td>0.1440</td>
</tr>
<tr>
<td>E1.PreTest-E1.DelayedTest</td>
<td>-0.3103</td>
<td>1.3406</td>
<td>56</td>
<td>0.1855</td>
<td>0.3709</td>
</tr>
<tr>
<td>E1.PostTest-E1.DelayedTest</td>
<td>0.0690</td>
<td>0.3752</td>
<td>56</td>
<td>0.7089</td>
<td>0.7089</td>
</tr>
</tbody>
</table>

Simple Effects for "GROUP x test" interaction is conducted to explain the interaction between group and test comparisons. Mendoza's Multisample Sphericity Test and Epsilons for GROUP X test was examined to check for any *Sphericity* violation (Table 6). The result showed that there is a violation found due to the small number of samples for test at Control (p-value = 0.0025) and test at Experiment (p-value = 0.0069), however, correction have been made by using GG, HF and CM. After the correction, simple effects for "GROUP x test" interaction was conducted (Table 7). The result showed that: (1) No significant difference found for the initial test between control and experiment group (p-value = 0.7624). (2) No significant difference found for the pre-test between control and experiment group (p-value = 0.7539). (3) Significance difference found for the post-test between control and experiment group (p-value = 0.0154; p < .05). (4) Significance difference found for the delayed test between control and experiment group (p-value = 0.0541; p < .10). (5) No significant difference found for all test in control group (p-value = 0.6465). (6) Significance difference found for all test in experiment group (p-value = 0.0000; p < .001).

Table 6. *Mendoza's Multisample Sphericity Test and Epsilons for GROUP X test*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Lambda</th>
<th>Approx. Chi</th>
<th>df</th>
<th>p-value</th>
<th>LB</th>
<th>GG</th>
<th>HF</th>
<th>CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>test at Control</td>
<td>0.0001</td>
<td>18.3672</td>
<td>5</td>
<td>0.0025 **</td>
<td>0.3333</td>
<td>0.7321</td>
<td>0.7970</td>
<td>0.7852</td>
</tr>
<tr>
<td>test at Experiment</td>
<td>0.0002</td>
<td>15.9899</td>
<td>5</td>
<td>0.0069 **</td>
<td>0.3333</td>
<td>0.7278</td>
<td>0.7917</td>
<td>0.7800</td>
</tr>
</tbody>
</table>

Table 7. *Simple Effects for "GROUP x test" interaction*
Multiple Comparison for "test at Experiment" using Holm's Sequentially Rejective Bonferroni Procedure was conducted to break down the result found in the simple effect for "GROUP x test" (Table 8). The factor is analyzed as dependent means and the alpha level used is 0.05. The result showed that: (1) initial test score is less than post-test score. (2) initial test score is less than delayed test score. (3) pre-test score is less than post-test score. (4) pre-test score is less than delayed test score. (5) initial test score is equal to pre-test score. (6) post-test score is equal to delayed test score.

Table 8. MULTIPLE COMPARISON for "test at Experiment" using Holm’s Sequentially Rejective Bonferroni Procedure

<table>
<thead>
<tr>
<th>Pair</th>
<th>Diff</th>
<th>t-value</th>
<th>Df</th>
<th>p</th>
<th>Adj.p</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1.InitTest-E1.PostTest</td>
<td>-1.4138</td>
<td>4.5901</td>
<td>28</td>
<td>0.0001</td>
<td>0.0046 E1.InitTest &lt; E1.PostTest *</td>
</tr>
<tr>
<td>E1.InitTest-E1.DelayedTest</td>
<td>-1.3448</td>
<td>4.0271</td>
<td>28</td>
<td>0.0004</td>
<td>0.0589 E1.InitTest &lt; E1.DelayedTest *</td>
</tr>
<tr>
<td>E1.PreTest-E1.PostTest</td>
<td>-1.0000</td>
<td>3.8079</td>
<td>28</td>
<td>0.0007</td>
<td>0.0914 E1.PreTest &lt; E1.PostTest *</td>
</tr>
<tr>
<td>E1.InitTest-E1.PreTest</td>
<td>-0.9310</td>
<td>3.0870</td>
<td>28</td>
<td>0.0045</td>
<td>0.1440 E1.InitTest &lt; E1.PreTest</td>
</tr>
<tr>
<td>E1.PreTest-E1.DelayedTest</td>
<td>-0.4138</td>
<td>1.6506</td>
<td>28</td>
<td>0.1100</td>
<td>0.3709 E1.InitTest = E1.PreTest</td>
</tr>
<tr>
<td>E1.PostTest-E1.DelayedTest</td>
<td>-0.0690</td>
<td>0.3864</td>
<td>28</td>
<td>0.7089</td>
<td>0.7021 E1.PostTest = E1.DelayedTest</td>
</tr>
</tbody>
</table>

+p < .10, *p < .05, **p < .01, ***p < .001

6. Discussion and Conclusion

Viat-map application allows students to build a kind of logical structure following the Toulmin Arguments concept. This kind of action will also form a kind of cognitive structure in their perception. By representing students with the visible logical structure in the application, they will also be able to view their own cognitive structure. Viat-map application also directs the students by following the teacher’s understanding of the text. Closed-ended approached is beneficial for directing the student to have the same level of understanding as the teacher and will be easier for the teacher to compare the mistakes made by the students. The map reconstruction approach has been implemented for the concept map (Hirashima, 2019) and several investigations have reported that it is useful to realize immediate and accurate diagnosis and feedback (Andoko et al., 2020; Pailai et al., 2017).

Based on the result above, the main objective of this study is to find out the comparison of students’ understanding and memory retention between students who use Viat-map and summarization conventional method. Firstly, to measure student understanding, the test score
was used as the representation of students understanding. Secondly, to measure memory retention, the delay-test score was used to measure the students’ memorization.

The first aim is to measure students’ understanding by using pre-test and post-test scores. Based on the calculation in, the experiment group having an improvement from pre-test to post-test (Table 8), while the control group remain the same (Table 7) where test at control found no significant difference (p-value = 0.6465). The comparison of post-test between experiment and control group also shown that experiment group overachieved the control group (Table 7; P-value = 0.0154; p < .05). The improvement from pre-test to post-test in experiment group also indicates that there is an improvement of score before using the application (pre-test) compared to the score after using the application (post-test). The experimental result indicated that the use of application helped the students to get a better test score than one that didn’t use. Assisting students in learning process by providing the require components to be used by the students could improve their understanding (Hirashima, 2019; Hirashima et al., 2015).

The second aim for memory retention was using the post-test and delay-test score. Based on the result, both groups can get the same score between post-test and delay-test score. However, considering that the delay-test score between control and experiment groups are different, where experiment group overachieved the control group (Table 7; p-value = 0.0541 +; p < .10) we can assume that the experiment group is better than the control group for memory retention. To show a better view of the overall test, a line chart of mean for each test in both groups is created (Figure 4). The differences of mean between groups for all tests are clearly showed. Other perspective of the score distribution for each test in both groups, box plot chart was also conducted (Figure 5). The initial test from both group shows that the score distribution is a little unbalance where the control group is denser than the experiment group. However, the score distribution was changed for pre-test where they are almost having the same distribution and means. A huge change occurred for post-test where means value for experiment group is higher than control group, however the score distribution is similar. Other changed is also occurred for delay-test where the distribution score for control group become wider compared to their post-test even the means remain the same.
In summary, constructing Toulmin arguments as logical structure offering students to understand the reading material in a different way. Using applications for study could also assist them to enrich their learning process. This study found that students who used Viatmap application have a better understanding compared to one who is not using the application. They also could retain their understanding 1 week after the lesson. However, there are several limitations for this study. First, this study only involves students from one University and one department that cannot represent the university students. Second, the number of sample size is only fifty-eight students which is very small. This study requires more evidence and many students to represent an ideal used of application in a classroom situation. For future research, it is important to conduct a comprehensive log analysis to give a better explanation and new points of view to improve the result of this study.

Acknowledgements

We would like to thank State Polytechnic of Malang for funding this research under Research grant number: SP DIPA-023.18.2.677606/2023.

References


Eftekhari, M., & Sotoudehnama, E. (2018). Effectiveness of computer-assisted argument mapping for comprehension, recall, and retention. ReCALL. https://doi.org/10.1017/S0958344017000337


Effects of a Machine Learning-empowered Chinese Character Handwriting Learning Tool on Rectifying Legible Writing in Young Children: A Pilot Study

Lung-Hsiang WONGa*, Guat Poh AWa, He SUNa, Ching-Chiuan YENb, Chor Guan TEOb, & Yun WENa

aNational Institute of Education, Nanyang Technological University, Singapore
bNUS-KEIO CUTE Center, National University of Singapore, Singapore
*lhwong.acad@gmail.com

Abstract: The logographic nature of Chinese script is a major dissuading factor for learning handwriting. The challenge is the complex psycholinguistic process behind handwriting. Thus, we developed AI-Strokes, a Chinese handwriting learning tool that assists teachers in facilitating students’ handwriting practice in various modalities, and provides personalized feedback for the students. By leveraging a trainable Machine Learning back-end framework, the tool diagnoses and scores students’ handwriting errors. This paper reports a pilot study in a Singapore primary school with an early prototype of AI-Strokes. Two classes of students went through AI-Strokes-based Chinese handwriting lessons (the experimental group) and conventional lessons (the control group) respectively. Pre- and post-tests were administered, and their handwriting processes were analyzed regarding errors in stroke orders, extra/missing strokes, and errors in stroke directions. The results show that the experimental group has yielded significantly better learning gains than the control group. It is posited that the personalized feedback of AI-Strokes has formed a feedback loop to support students’ trial-and-error process in improving their handwriting skills. The multimodal handwriting task design may have also fostered their orthographic awareness through the activation of alternative psycholinguistic pathways during their handwriting lessons.

Keywords: AI in education, Computer-assisted language learning, Learning of Chinese handwriting, machine learning

1. Introduction

The logographic nature of Chinese script is a major dissuading factor for novice learners in learning handwriting. Chinese handwriting involves complex visual-perceptual-motoring processing (Haas & Rees, 2010). Acquiring such an integrated processing skill is cognitively demanding for, in particular, young Chinese as a second language learners including most young students in Singapore (Wong et al., 2011). Furthermore, in the context of Singapore primary schools, hampered by big class sizes and limited lesson time, timely formative feedback by the teachers during individual students’ learning process is hardly materialized.

This study seeks to address the challenges by developing AI-Strokes, a web-based Chinese handwriting learning tool that assists teachers in facilitating handwriting lessons with various types of prompts, and providing personalized feedback to the students. By leveraging a trainable Machine Learning back-end framework, AI-Strokes could automatically diagnose and score handwriting errors of individual students.

This paper reports on a pilot study in a Singapore school with an early prototype of AI-Strokes. The quasi-experimental study involved two Primary 2 classes who went through AI-Strokes-based Chinese handwriting lessons (the experimental group) and conventional lessons (the control group) respectively. Pre- and post-tests were administered to both
groups, and their handwriting processes were analyzed in terms of errors in stroke orders, extra/missing strokes, and errors in stroke directions. The research question is as follows, “Would the incorporation of AI-Strokes into the Chinese as a second language handwriting lessons in lower primary school levels help reduce students’ Chinese handwriting errors, as compared to conventional lessons without personalized ICT support?”

2. Literature Review

2.1 Chinese character handwriting

The Chinese scripts are a principled and rule-based system. Each Chinese character comprises one or more components spatially arranged with certain principles (Liang, 2004). Each character is composed of strokes, the basic and smallest unit of a character without any semantic connotation (Lam & McBride, 2018), with fixed, codified stroke orders for identical components in different characters. Handwriting processes involve multifactorial pathways that connect phonology, orthography, and semantics (Yin et al., 2005). The process of writing a Chinese character can be summarized into three steps as follows,

- **Step 1**: Retrieving the orthographic representation (or, the mental image) of a previously learnt character from long-term memory and store in working memory – specifically, the speech sound of the character activates the phonological representation of the character, which activates the lexical representation related by meaning via the lexical semantic pathway (Lam & McBride, 2018).

- **Step 2**: Processing of the mental image in working memory with the stroke orders.

- **Step 3**: Actual handwriting actions which involve the following sensorimotor performance components, all activated almost simultaneously (Klein, et al., 2011):- visual perception, visual skills, fine motor, and visual motor.

There are at least two levels of objectives for learning to handwrite Chinese characters: "legible writing" and "orthographic retrieval". "Legible writing" (Tsai et al., 2012) focus on Steps 2 and 3 - to handwrite “legible” characters with correct strokes, stroke orders, spacing, etc. Typical learning designs are “copying tasks”, e.g., to display a full character for students to copy. Conversely, "orthographic retrieval" (Qu & Damian, 2015) tackles Step 1 where students handwrite characters without visual references of “model characters”. The activities could be, (a) written picture naming task (e.g., to show an image of a tree, and the student writes 树); (b) dictation task (e.g., a sound clip utters, “生气的 生”, and the student writes 生); (c) translation task (for second language students; e.g., to show the English word “flower” and the student translates it mentally and write 花).

This study focuses on assisting teachers in improving young students’ competencies in performing Steps 1 and 2 of the Chinese character handwriting process with AI-Strokes.

2.2 Automation of diagnosis of students’ handwriting

Traditional technology has been used to facilitate the recognition and scoring of Chinese handwriting (Hsiao et al., 2015) according to established systems and rules. Using neural networks for recognition, compared to classical methods, a system may present an advanced alternative to human-based scoring. The ability to learn and establish parameters during the training process means no manual hardcoding is needed. This makes the system scalable and sustainable, and resilient to changes in handwriting styles.

The use of machine learning for recognizing handwritten numbers and English characters has been explored to a great extent, with early works on neural networks recognizing numbers with an error rate of 0.7% (LeCun et al., 1998). The use of Convolutional Neural Networks for recognizing handwritten English characters peaked in 2011, when Ciresan and team achieved an error rate of just 0.27% (Ciresan et al., 2011), making it comparable to human-like performance.
Diagnosis of Chinese handwriting, however, has not been explored in depth, with some having moderate success at handwriting recognition (Bai et al., 2014; Zhang et al., 2019; Zou et al., 2019). These attempts are empowered by large datasets or corpuses of written characters, whilst the system being developed performs machine learning on the process of the writing itself, thereby giving it more data to work with while doing diagnosis. AI-Strokes can also point out the errors made during the writing process (e.g., wrong directions of strokes), which the systems reported in the aforementioned publications cannot.

3. System Design of AI-Strokes

AI-Strokes is a web-based teaching aid that helps teachers manage and record the process of each student's Chinese handwriting and provide personalized diagnosis on their strengths and weaknesses, empowered by the advanced Machine Learning backend. With more student handwriting data being collected, the system can self-learn to analyze and score handwritten characters that it has not been trained with. The web-based tool should be accessed by touchscreen devices for a more natural handwriting experience. The system design is intended for use in classroom lessons (e.g., learning new characters in the textbook). The system user interface (UI) comprises teacher’s console and students’ client.

Before each lesson, the teacher creates a new lesson plan with the teacher’s console UI via a laptop or a tablet computer by specifying the characters to practice. At the lesson, the teacher selects the pre-stored lesson plan and launches it, with the teacher’s console UI being projected to the screen. Each student who is assigned a tablet may then log in and join the lesson with the student UI. Every time when the teacher administers a character for the entire class to practice, (s)he may provide the prompt through teacher’s console in one of the following modalities: (1) the character itself (copying task); (2) an image or an animation that depicts the meaning of the character (picture naming task); (3) the equivalent English word (translation task); (4) a sound clip that pronounces the character with disambiguation (dictation task). The copying task corresponds to “legible writing” while the three other tasks correspond to “orthographic awareness” (see section 2.1).

In both modes, students are given immediate feedback (Figure 1) upon submission of their handwriting via the automated scoring system powered by the Machine Learning backend. The Machine Learning backend scores each student's writing with how well it matches the correct answer in terms of number of strokes, order of strokes, directions of strokes, and speed. The aspects of handwriting to be diagnosed is a subset of the validated Tseng’s Handwriting Problem Checklist (THPC) (Tseng, 1993). According to THPC, there are 24 common Chinese character handwriting problems among young students. Only 16 of them may be automatically diagnosed by typical AI-empowered web-based software. For this early prototype of AI-Strokes, automated diagnoses on 8 of the problems are implemented, lumped into 4 categories as listed in the score card in Figure 1. In this early prototype, a submitted handwritten character may be awarded a star for each of the achievements of correct number of strokes, order of strokes and directions of strokes. Thus, the full score is 3 stars.

In teacher’s console, there is an additional classroom view. After administering a character for the students to practice, the teacher may switch to the classroom view with all students' writing are displayed in real-time. The teacher may get a glimpse of how each student has done, or project it on the screen for class wide comparison and discussion.

4. Research Design of the School-based Pilot Study

Fifty (50) Primary 2 students from a public school in Singapore participated in the study, including 28 from the experimental group/class and 22 from the control group/class. The experiment comprised four one-hour lessons that spanned across 1.5 months. During the experiment, the teachers of both classes tapped on the regular Chinese lessons to conduct handwriting activities with the same sets of characters. However, AI-Strokes was employed in the experimental class while the prevailing handwriting instructions were carried out in the
control class. In the control class, the teacher first demonstrated writing a character on the whiteboard. The students then practiced writing individual characters in the air with fingers or on their mini whiteboards and displayed them to the teacher. Eventually, all students were drilled in handwriting in their exercise books by copying the given characters or by dictation.

The pre-test and the post-test took place before the first lesson and after the final lesson respectively and were administered as copying tasks. At each test, students used Al-Strokes to handwrite 10 characters – 瓜, 亲, 哭, 温, 爱, 起, 笔, 看, 极. The system captured the students’ writing processes (stroke orders) and products (the final appearance of each character) for analysis. The characters were selected from 30 characters covered in the four lessons with a maximum variety in the character components.

Stroke-by-stroke images of the writing process of each character were captured by the system for analysis. The process was assessed in: (a) Errors in stroke orders (“s-orders” hereafter): The number of strokes written in wrong orders; (b) Number of extra/omitted strokes (“eo-strokes” hereafter): The sum of the absolute number of extra strokes and the absolute number of omitted strokes – including cases where the student breaks one stroke into two or more strokes (e.g., writing the third stroke of 瓜 in two strokes – counted as 1 extra stroke), or write two or more strokes in one stroke (e.g., writing the three-stroke component 亻 in one stroke or two strokes – counted as 2 or 1 omitted stroke(s)); (c) Errors in stroke directions (“s-dir” hereafter) (e.g., writing 水 (a horizontal stroke from left to right) as 亻 (a tilted stroke from right to lower left) is considered an error). Thus, for all three indicators, the lower the numbers are, the better the student’s performance is.

5. Findings

We employed SPSS 28 to analyze the pre- and post-test data (Table 1). To determine whether parametric or non-parametric Analysis of Covariance (ANCOVA) should be performed on the three indicators respectively (Corder & Foreman, 2009), the normal distribution tests of Shapiro-Wilk were executed. The results showed that there were normal distributions in both the pre- and post-test scores of s-orders (p=.324>.05 and p=.258>.05 respectively), while there was normal distribution in the pre-test score of eo-strokes (p=.153>.05) but no normal distribution in the post-test score of the same indicator (p<.05), and there were no normal distribution in both the pre- and post-test scores of s-dir (p<.001 and p<.001 respectively). Therefore, parametric ANCOVA was performed to compare the pre- and post-test scores of s-orders between the two groups, while Quade’s non-parametric ANCOVA was executed to compare the eo-strokes and s-dir scores respectively.

The results showed that the handwriting performance of the experimental group was improved significantly as compared to the control group. By comparing the pre- and post-test
means of the control group, there were even dips in their performance in s-orders and eo-strokes. Notwithstanding, the dips are insignificant, according to our supplementary paired samples t-tests (s-orders: p=.079>.05; eo-strokes: p=.353>.05; s-dir: p=.164>.05). On the contrary, the experimental group exhibited significant improvements in all indicators, as seen in the results of paired-sample t-tests (s-orders: p<.001; eo-strokes: p<.001; s-dir: p<.05).

Table 1. Summary of descriptive statistics and results of parametric or non-parametric ANCOVA on the three indicators of students’ handwriting performance

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Group</th>
<th>Pre-test Mean (SD)</th>
<th>Post-test Mean (SD)</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-orders (parametric ANCOVA)</td>
<td>Experimental</td>
<td>77.43 (4.23)</td>
<td>70.93 (4.94)</td>
<td>11.447</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>73.86 (3.91)</td>
<td>75.77 (3.89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eo-strokes (Quade’s ANCOVA)</td>
<td>Experimental</td>
<td>3.96 (2.43)</td>
<td>1.96 (2.73)</td>
<td>28.587</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>5.55 (3.14)</td>
<td>5.86 (1.71)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s-dir (Quade’s ANCOVA)</td>
<td>Experimental</td>
<td>1.75 (1.84)</td>
<td>0.89 (1.69)</td>
<td>8.812</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>2.14 (2.05)</td>
<td>1.55 (1.54)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Levene’s Test of Equality of Error Variances: F=.252, p=.618>.05.

High error rates of stroke orders in both tests and across both groups are observed in the statistics. The correct total number of strokes of the 10 tested characters is 88. Thus, the error rates according to the four mean scores of s-orders is ranging from 80.5% to 88.0%. The roots of the phenomenon can probably be found in the student demography and the teachers’ instructional approaches, which will be discussed in the next section.

6. Discussion and Conclusion

This study found that the experimental group which experienced the AI-Strokes-empowered Chinese handwriting lessons has yielded significantly better learning gains in the three measured indicators, as compared to the control group. It is posited that the personalized feedback feature of AI-Strokes has formed a feedback loop to trigger and support students’ trial-and-error process to improve their handwriting skills. The multimodal handwriting tasks have also played the role of fostering students’ orthographic awareness through activation of alternative cognitive and psycholinguistic pathways during their handwriting lessons.

Yet why were there high error rates in stroke orders even at the post-test? Scholars (Lam & McBride, 2018; Law et al., 1998) reiterated the need to teach general stroke-order rules. When students have internalized these rules, they could handwrite any novel character with correct order. However, during the limited lesson time, teachers tend to focus on demonstrating stroke orders of individual characters, and only introduce scattered stroke order rules by chance. Students might ignore these rules but just memorize the stroke order of each character. In paper-based assignments, the correctness of stroke orders cannot be assessed as only the completed handwriting is submitted for grading. Thus, most students are not motivated to learn correct stroke orders as long as their “final products” “look right”.

In our study, however, the experimental group students have constantly received feedback from AI-Strokes after handwriting each character. If the feedback indicated that their stroke order was not correct, they might try to figure out what went wrong by recalling what the teacher had taught before or asking around their classmates, with the hope of scoring higher in their next writing attempt. Thus, we postulate that AI-Strokes had subtly influenced the experimental group students to start paying attention to the stroke orders.

Despite the significant improvement in stroke orders among the experimental group students, their error rates in the post-test are still relatively high. By comparing their stroke orders between the pre- and post-tests, it is observed that at the post-test they were more likely to handwrite the characters with partially correct stroke orders. Yet most of the time they did not manage to handwrite whole characters with perfectly correct stroke orders. The reasons are two-folded. First, the prevailing issue of the lack of systematic explicit teaching of stroke order rules would have limited the students’ development of stroke order
“competency”. Second, the feedback given by the current AI-Stroke prototype is relatively coarse-grained and is unable to pinpoint the sources of errors; but just alerts the students that "something was wrong" and they need to find out the errors through other means.

To address these limitations, we advocate a holistic revamp in AI-Stroke-based lessons. In the instructional aspect, the teachers should set aside the time to systematically cover the general stroke order rules. When a new rule is introduced, the teachers should let students practice writing multiple characters where this rule is applicable (rather than insisting to practice the characters appearing on the current textbook passage) on AI-Stroke, to help them internalize the rule and later apply it to writing novel characters with the same “stroke configuration”. In the technical aspect, AI-Stroke should be upgraded to offer finer-grained feedback. AI-Stroke may play a crucial role in providing personalized formative assessment of handwriting, which is unable to be performed by the teachers.

Given its current machine learning backbone engine, AI-Stroke has the potential to be further upgraded to incorporate the capability of diagnosing additional aspects of THPC including spacing, spatial relationship, size and formation. More young children’s handwriting samples need to be collected as training data for this purpose. With an upgraded AI-Stroke system, a longer-term study can be carried out with a full-fledged Chinese handwriting lessons to develop a stronger psycholinguistic and handwriting foundations in the students.

References


Using TAASSC to Investigate Fine-Grained Grammatical Complexity in Reading Texts of Two High-Stakes English Tests in China

Shengshu LIN
School of Foreign Languages, Nanjing University of Posts and Telecommunications, China
shengshu2016@hotmail.com

Abstract: Fine-grained grammatical complexity measures are better predictors than large-grained indices of register variation and writing of different levels in that they provide in-depth explanation of what accounts for syntactic complexity. Computer technologies have made it possible to study these fine-grained measures based on large corpora. Using the Tool for Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) (Kyle, 2016), the present study investigates phrasal and clausal complexity in reading texts of two high-stakes English tests in China (CET-4 and NETEM) different in difficulty level. The results showed that clausal and phrasal complexity indices could predict the two tests. More importantly, in combined analysis of the two types of measures, the results demonstrated that noun phrases, particularly those with phrasal modifiers, were better predictors of the more difficult test NETEM. These findings support the importance of noun phrases and the diverse modifiers they take in informational texts. Findings of this research will help college students advance their English learning with a specific purpose and in a right direction.

Keywords: clausal complexity, phrasal complexity, reading texts, tests

1. Introduction

Reading comprehension has always been an important task in time-limited English tests. Yet grammatical complexity is given scant attention in reading accuracy despite its wide attention in writing. Grammatical complexity in spoken and written English and its historical changes were fully documented in Biber and Gray (2016). Recent research showed that fine-grained grammatical complexity measures were more effective in accounting for sentence length and writing quality (Kyle & Crossley, 2018; Zhang & Lu, 2022; Zhang & Liu, 2022; Gray, Geluso, & Nguyen, 2019). According to Biber, Gray and Poonpon (2011), advanced phrasal features are not acquired naturally and not always successfully acquired. The empirical research by Priven (2020) showed that fine-grained phrasal complexity pervasive in university textbooks caused difficulty to ESL readers. In Bychkovska (2021), the improved quality of writing by ESL writers after explicit teaching of phrasal features also lent support to the importance of fine-grained complexity features in reading. Due to the washback effects of tests, reading texts of a test are a critical source of language input for learners. For example, Chinese learners in preparation for NETEM (the National Entrance Test of English for MA/MS) use past years’ real tests as major learning materials. The present study is conducted with the purpose of shedding light on the fine-grained linguistic features in reading texts of two important tests in China and helping test takers advance their learning in a right direction.

2. Research Background

2.1 Fine-Grained Measures
Measures of grammatical complexity have undergone dramatic changes with mounting attention to fine-grained measures because they can account for what contributes to the length of length-/ratio-based large-grained measures.

In this line of research, Biber and colleagues have always employed form-function framework to define linguistic features. They believe only by taking structural types and their grammatical functions into consideration can register variations be fully described, because the same structural type may have different functions in different registers. By now, the eight form-function categories have included over 130 specific linguistic features and proved fruitful. Register studies showed that academic discourse was complex in the dense use of phrasal features as nominal modifiers, while spoken conversation was clausally complex. Language learners progress along a sequence of linguistic features acquiring grammar in conversation first and that of writing later (Biber et al., 2011). Their theory was supported by many researchers (Staples et al., 2016; Parkinson & Musgrave, 2014; Gray et al., 2019).

Although fine-grained measures based on form-function framework are easy to understand, the tools used to tag and count linguistic features are not publicly available. From the usage-based theory of language acquisition, Kyle (2016) developed the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) by utilizing Stanford Neural Network Dependency Parser (version 3.5.1), making it possible to analyze a large number of fine-grained linguistic units. Research conducted with TAASSC showed that fine-grained measures were better predictors than large-grained measures of learner writing (Kyle & Crossley, 2018; Zhang & Liu, 2022; Zhang & Lu, 2022). Kyle and Crossley (2018) was the first research that directly compared the predictive power of large-grained indices (calculated by L2SCA), clausal and phrasal complexity in TOEFL independent writing by L2 writers from different language backgrounds. Large-grained indices explained 5.8% of variance in holistic essay scores while fine-grained clausal complexity explained 3.0% and fine-grained phrasal complexity explained 18.9%. The large predictive power was supported by recent studies (Zhang & Liu, 2022; Zhang & Lu, 2022). Zhang and Lu (2022) examined Chinese university students’ English expository writing. Their results showed that fine-grained indices accounted for 22.2% of the variance in students’ writing quality, significantly higher than 7.2% by large-grained indices. Zhang and Lu (2022) investigated grammatical complexity in application letters and argumentative essays by Chinese students. Traditional indices explained 20.2% of the variance in letter scores, and 15.7% of the variance in essay scores while fined-grained measures explained 31.9% and 30.6% respectively.

The research aforementioned focused on the predictive power of fine-grained measures in writing while investigation of reading texts in tests has been ignored by now. Some fined-grained measures such as (multi-)noun as premodifiers and postmodifying prepositional phrases are compressed with dense information, and may cause confusion in reading (Priven, 2020). Therefore, utilizing TAASSC, the present study makes contributions to this area of research by investigating the reading texts of two important English tests in China.

2.2 CET-4 and NETEM in China

College English Test Band 4 (CET-4) as a proficiency test for university students in China is influential among college students in China. In the third year or even earlier, millions of college students will start to prepare for NETEM held in December annually to pursue postgraduate studies. Reading tasks common in CET-4 and NETEM require test takers to answer five multiple-choice questions after reading a passage. Although reading texts in these two tests are extracts from some publications in English-speaking countries including The Economist, The Guardian or other academic reading materials, CET-4 is easier for its simpler vocabulary, shorter sentence length and text length. Since NETEM is regarded as the most challenging test for non-English majors, some publications encourage test takers to use complex clauses to obtain higher scores without mentioning phrasal features (Zhou, 2022). The status quo reflects the stereotype of academic writing as clausally complex and clausal elaboration as a focus in most EAP teaching (Parkinson, 2020). To reveal grammatical complexity in the reading texts of the two tests, the present study addresses three research questions:
RQ1. What clausal complexity measures can predict CET-4 texts and NETEM texts?
RQ2. What phrasal complexity measures can predict CET-4 texts and NETEM texts?
RQ3. What clausal and phrasal complexity measures when considered together can predict CET-4 texts and NETEM texts?

3. Methodology

3.1 Corpus

Two corpora (NETEM and CET-4) used in this study were built by the author. NETEM is held annually with four reading passages in each real test paper of the previous 13 years from 2011 to 2023. 52 texts were collected with a total of 23,023 words. CET-4 reading texts were more easily available because it is administered twice a year with 6 different reading passages each time. In order to control the effect of time on CET-4 language use, we randomly selected four texts in each year to ensure that CET-4 reading passages are evenly distributed from 2010 to 2022. 52 CET-4 texts totaled 18,411 words. All texts were saved in a .txt format with the filename "year_test_text_number" that can be processed by TAASSC.

3.2 Tools and Procedures

The current study uses TAASSC (Kyle, 2016; Kyle & Crossley, 2018) to investigate clausal and phrasal complexity. From usage-based theory perspective, the clausal complexity in TAASSC refers to a main verb and its associated structures. TAASSC creates a dependency representation of each sentence, that is, the functional categories and their grammatical relations. It includes 31 clausal indices, of which 2 are general indices of clausal complexity and 29 calculate the average number of particular units per clause. The phrasal complexity involves 7 noun phrase types and 10 dependent types. Phrasal measures produced by TAASSC have pronoun version and no-pronoun version, each including 66 indices. This study used no-pronoun version because pronoun-headed phrases are not common (Lan et al., 2019). Of the 66 phrasal indices, 16 calculate the average number of dependents per each phrase type and for all phrase types, and their standard deviations (SDs). 10 indices calculate the occurrence of particular dependent types regardless of the type of noun phrases they occur in. 40 indices calculates the average occurrence of particular dependent types in particular types of noun phrases. A higher value indicates a higher frequency of a linguistic unit. SDs are used to measure syntactic variation.

These 97 complexity indices collected by TAASSC were entered in SPSS as independent variables. Since rareness of some linguistic structures is inherent in fine-grained analysis (Kyle & Crossly, 2018), we removed those that displayed ‘0’ value. Due to nonnormal distributions of the variables, non-parametric independent samples Mann-Whitney U test was performed to identify the indices on which CET-4 and NETEM differed significantly (p < 0.5). This step identified the independent variables that were used for binary logistic regression. Binary logistic regression was performed on the significantly different clausal complexity indices and phrasal complexity indices respectively to determine which were predictors of reading texts of the two tests. The results from these two steps answered RQ1 and RQ2. To answer RQ3, a combined analysis of the significantly different clausal and phrasal indices was conducted to determine which indices entered the final regression model.

4. Results and Discussion

Table1, Table 2 and Table 3 include the variables that entered into binary logistic regression. The variables with asterisks were predictors of texts in the two tests at the statistical level of
0.05. Exp(B) values more than 1 indicated the complexity indices had predictive effects on texts in NETEM, while Exp(B) values less than one predicted CET-4 texts.

Table 1. Logistic Regression of Clausal Complexity

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.652</td>
<td>5.262</td>
<td>0.482</td>
<td>0.488</td>
<td>0.026</td>
</tr>
<tr>
<td>2</td>
<td>6.568</td>
<td>4.955</td>
<td>1.757</td>
<td>0.185</td>
<td>712.143</td>
</tr>
<tr>
<td>3</td>
<td>4.534</td>
<td>2.592</td>
<td>3.060</td>
<td>0.080</td>
<td>93.152</td>
</tr>
<tr>
<td>4</td>
<td>48.906</td>
<td>19.923</td>
<td>6.026</td>
<td>0.014*</td>
<td>1.736</td>
</tr>
<tr>
<td>5</td>
<td>7.367</td>
<td>3.322</td>
<td>4.919</td>
<td>0.027*</td>
<td>1583.482</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.215</td>
<td>1.549</td>
<td>7.401</td>
<td>0.007</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Table 2. Logistic Regression of Phrasal Complexity

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.645</td>
<td>2.685</td>
<td>0.375</td>
<td>0.540</td>
<td>0.193</td>
</tr>
<tr>
<td>2</td>
<td>3.254</td>
<td>2.047</td>
<td>2.527</td>
<td>0.112</td>
<td>25.882</td>
</tr>
<tr>
<td>3</td>
<td>3.965</td>
<td>1.572</td>
<td>6.363</td>
<td>0.012*</td>
<td>52.720</td>
</tr>
<tr>
<td>4</td>
<td>0.678</td>
<td>0.547</td>
<td>1.538</td>
<td>0.215</td>
<td>1.970</td>
</tr>
<tr>
<td>5</td>
<td>1.191</td>
<td>0.649</td>
<td>3.369</td>
<td>0.066</td>
<td>3.290</td>
</tr>
<tr>
<td>6</td>
<td>-31.467</td>
<td>13.879</td>
<td>5.141</td>
<td>0.023*</td>
<td>147.287</td>
</tr>
<tr>
<td>7</td>
<td>4.992</td>
<td>4.275</td>
<td>1.364</td>
<td>0.243</td>
<td>5071.844</td>
</tr>
<tr>
<td>8</td>
<td>8.531</td>
<td>4.518</td>
<td>0.159</td>
<td>0.690</td>
<td>6.054</td>
</tr>
<tr>
<td>Constant</td>
<td>-7.178</td>
<td>2.487</td>
<td>8.329</td>
<td>0.004</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 3. Logistic Regression of Clausal Complexity and Phrasal Complexity Combined

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.666</td>
<td>2.685</td>
<td>0.162</td>
<td>0.688</td>
<td>0.070</td>
</tr>
<tr>
<td>2</td>
<td>3.663</td>
<td>3.374</td>
<td>0.330</td>
<td>0.565</td>
<td>38.992</td>
</tr>
<tr>
<td>3</td>
<td>5.389</td>
<td>3.179</td>
<td>2.873</td>
<td>0.090</td>
<td>218.883</td>
</tr>
<tr>
<td>4</td>
<td>78.275</td>
<td>28.267</td>
<td>7.668</td>
<td>0.006*</td>
<td>9.873</td>
</tr>
<tr>
<td>5</td>
<td>3.541</td>
<td>4.187</td>
<td>0.715</td>
<td>0.398</td>
<td>34.494</td>
</tr>
<tr>
<td>Constant</td>
<td>-10.874</td>
<td>3.435</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To answer RQ1, Table 1 shows that bare noun phrase temporal modifiers and auxiliaries are more common in NETEM texts, and the more use of these features indicates the higher probability that reading texts are from NETME. Many of NETME texts are extracted from top newspapers, where temporal words are important linguistic resources in narrating events. The ability to capture the temporal dimension is essential for understanding news events. Auxiliaries can convey much grammatical information such as aspect, tense, voice and modality. More importantly, modal auxiliaries express a variety of meanings related to concepts such as ability, permission, necessity, and obligation (Biber et al., 1999). It is very common in NETEM reading comprehension that questions are asked about attitudes of the news writer or people involved in news events. NETME is more complicated than CET-4 in the use of temporal words to connect events and auxiliaries to express grammatical information and attitudes, though both tests use news extracts as a major source of reading texts.
With regard to RQ2, Table 2 shows that NETEM employs diverse modifying structures in noun phrases as a subject, as is illustrated by Excerpt 1. Excerpt 1 includes adjective “personal”, (multi-)noun “grievance” and “Employment Relations”, of-preposition “of New Zealand’s...2000”, and possessive “New Zealand’s”.

Excerpt 1: The personal grievance provisions of New Zealand’s Employment Relations Act 2000 (ERA) prevent an employer from firing an employee without good cause. (2022_NETEM_text_4)

A subject with various types of modifiers are quite common in NETEM reading texts because the extended subject packages dense information into phrasal premodifiers. Another index that makes reading texts more likely to be NETEM texts is the use of more possessives modifying a noun in direct object positions. Possessives, nouns and adjectives are the most compressed modifiers in noun phrases (Biber & Gray, 2016). However, using possessives, especially genitive ‘s is not encouraged among Chinese students because it may strengthen the stereotype of “Chinese English” since its counterpart structure in Chinese is prevalent. In this sense, our findings counter this stereotype and highlight the importance of these features in reading. According to Biber and Gray (2016), the use of verbal modifiers is viewed as more structurally elaborate rather than compressed, and is more common in spoken conversation. This is supported by our finding that more use of verbal modifiers of a noun will increase the likelihood of reading being CET-4 texts. Instead, phrasal modifiers are a common feature in passages with information focus (Biber & Gray, 2016).

Results in Table 3 provide answers to RQ3. As can be seen in Table 3, when clausal and phrasal complexity were combined, two new phrasal indices entered the regression model: dependents per object of prepositional phrases and dependents per agent (SD), while only bare noun phrase temporal modifiers per clause was retained as clausal index. Modifying devices of agents are also more diverse in NETEM reading. The index of dependents of prepositional objects further affirmed that NETEM were more likely to use extended noun phrases with embedded modification as illustrated by Excerpt 2. The prepositional phrase “of human activity” also takes an embedded modifier “in recent climate change”, which again takes an adjective “recent” and a noun “change” as its modifiers.

Excerpt 2: These would emphasise the primacy of human activity in recent climate change and encourage discussion of mitigation measures. (2023_NETEM_text_1)

Based on Biber et al.’s (2011) language development theory, extended prepositional phrases as post-modification are one of the most advanced features of academic language, and also the most difficult for both native and EFL language learners. NETEM is more likely to use diverse dependents to add information in compressed structures. According to Kyle and Crossley (2018), Zhang and Lu (2022), phrasal complexity is more predictive of higher level learner writing. This study further affirmed that, in English news, fine-grained indices, phrasal level features in particular, correlated significantly with the degree of difficulty and revealed what specific patterns contributed to difficulty of textual sentences in NETME. It is noteworthy that all the indices included in the regression model of combined analysis were realized linguistically by noun phrases that may take a variety of modifiers. This lends support to Biber and Gray’s (2016) claim about noun phrases that they are prevalent in information writing, especially those with phrasal modifiers. In view of empirical evidence that extended noun phrases may cause difficulty in reading university textbooks to L2 learners (Priven, 2020), noticing and understanding complex nominal modifiers should be emphasized in teaching advanced English learners. For those who prepare for NETEM or plan their learning after passing CET-4, using these compressed phrasal features to convey dense amounts of information should be a central area of learning.

5. Conclusion

Utilizing TAASSC, the present study provides insights into what fine-grained measures are predictive of reading texts of two important tests for university students in China. Findings will help millions of students focus on the linguistic features that are more typical of NETEM rather
than just spending time on clausal subordination. Future research can include vocabulary as a predictor variable, comprehensively looking at what vocabulary characteristics and linguistic features can predict reading texts in tests for different purposes and at different stages of English learning.

Acknowledgements

This study was supported by the 11th China Foreign Language Education Fund (ZGWJYJ11Z025) and Research on Philosophy and Social Sciences of Jiangsu Universities in 2022 (2022SJJYB0101).

References


Using Multidimensional Analysis to Investigate the Extrapolation Inference of a High-Stakes Test

Shengshu Lin
School of Foreign Languages, Nanjing University of Posts and Telecommunications, China
shengshu2016@hotmail.com

Abstract: The degree of correspondence of situational characteristics between test tasks and target domains is believed to affect how similar language use on tests is to the target domains (i.e., extrapolation). Multidimensional analysis (MDA) has gained in popularity in extrapolation research because it not only interprets language use functionally associated with situational characteristics but also provides a quantitative method to investigate co-occurring linguistic features. Using the automatic tool Multidimensional Analysis Tagger (MAT), the present paper applies MDA to explore the linguistic features between writing of the National Entrance Test of English for MA/MS (NETEM) and disciplinary writing on four dimensions in Biber’s (1988) seminal work. Overall, NETEM writing differed significantly from all four disciplinary domains on Dimension 1 and Dimension 4, and showed similarities with Humanities and Arts papers but differed from the other three domains on Dimension 2 and Dimension 3. NETEM writing was weak in its extrapolation to real academic contexts.

Keywords: Multidimensional analysis, writing, linguistic features

1. Introduction

Whether test tasks elicit linguistic features that students will use in real academic contexts is an important extrapolation concern in argument-based validity (Kane, 2012). Take TOEFL iBT for example. The inclusion of source-based integrated writing task in 2005 makes TOEFL Writing Section more representative of language use in real learning environments (Llosa, Grapin, & Friginal, 2019). The underlying assumption of extrapolation inferences is that similar situational characteristics of the target domain elicit the linguistic features associated with it (LaFlair & Staple, 2017; Kyle et al., 2021). The emphasis on communicative functions associated with situational characteristics in tests justifies the application of Multidimensional analysis (MDA) (Biber, 1988) to the investigation of extrapolation of test performances to those in the target domain. MDA takes the register perspective, believing that some linguistic features are more frequently used in a register because “they are functionally adapted to the communicative purposes and situational contexts of texts from that register” (Biber & Conrad, 2019, p. 2). More importantly, this method provides a corpus-based quantitative method to compare a large number of linguistic features with shared communicative functions (LaFlair & Staple, 2017; Yan & Staples, 2019). The present study is the first empirical research that employs this method to examine the linguistic features elicited in the writing task of a high-stakes test in China. The purpose of this study is twofold: 1) to demonstrate how MDA can be used to investigate the extrapolation inference of this test; 2) to discuss the implications for test developers and learners.

2. Research Background

Language users make systematic lexical and grammatical choices appropriate to the situational characteristics of the register (Biber, 2019). Utilizing computational techniques, MDA takes a bottom-up approach to study a large number of fine-grained lexico-grammatical
features distributed in a register and identify co-occurrence patterns with the same underlying functional underpinnings (Biber, 2019). The systematic co-occurring sets of linguistic features are factors or dimensions, and should be interpreted functionally with respect to situational characteristics. It is based on co-occurrence linguistic features that important register differences can be revealed and explained.

In recent years, MDA has attracted mounting attention in validation research, particularly in the investigation of extrapolation inferences because MDA not only extends TLU domain but serves as a quantitative framework for examining the linguistic relevance between test performances and the domains to which they are extrapolated (LaFlair & Staples, 2017; Staples et al., 2018). Weigle and Friginal (2015) was the first empirical research using additive MDA to examine the comparability of TOEFL Independent writing with disciplinary papers along four dimensions established by Hardy and Römer (2013). Results indicated that TOEFL Independent writing was characterized by more narrative expressions of opinion and stance, and expressions of possibility, whereas the latter relied more on linguistic features associated with procedural information. They concluded that Independent writing task was limited in its generalization to real academic contexts. These findings were supported by Llosa et al. (2019) who used the same four dimensions and extended their research by comparing TOEFL Integrated writing with disciplinary papers. Llosa et al. found that this writing task complemented Independent task by eliciting linguistic features associated with information and argumentation. Based on these findings, researchers claimed that TOEFL Writing Section approximated successful academic writing than Independent task alone. Utilizing the four dimensions by Gardner et al. (2019), Staples et al. (2018) compared TOEFL writing with the narrower registers such as case studies and argumentative essays in disciplinary writing. Again, the previous findings about Independent writing were supported. However, Staples et al. pointed out that although TOEFL Integrated writing was better representative of disciplinary writing, its overall patterns were more similar to Independent writing than they were to most disciplinary registers. As a result, the extrapolation of TOEFL writing was still a problem.

The current study contributes to this area of study by investigating Writing Task 2 of the National Entrance Test of English for MA/MS (referred to as NETEM writing throughout the paper) in China, a national high-stakes test that aims to select candidates with required English proficiency for future intellectual pursuits. NETEM with millions of test takers is regarded as the most challenging English test for non-English majors. NETEM writing is an essay of 160-200 words based on one cartoon or a set of cartoons. Test takers are required to describe the cartoon(s) briefly, interpret the meaning, and give comments. The differences of NETEM writing from disciplinary papers in situational characteristics including the topic, the purpose, and the production circumstances may result in different choices of linguistic patterns. Our research into the linguistic features in these two situations will inform language testing and teaching. The following research question will be addressed:

How similar are the linguistic features of NETEM writing to those of graduate level disciplinary papers?

3. Methodology

3.1 Corpus

The corpora for the current study consist of NETEM writing corpus and a reference corpus. The former contains samples of NETEM writing from the past ten years’ (2013-2022) real papers. These papers were compiled by ten different educational training and services providers and published by some well-known publishing companies in China. Although NETEM directions suggest 160-200 words, test takers are encouraged to produce over 180 words in order to score high. In the present research, the type-token ratio in MAT was set at 180 and four essays less than 180 words were excluded. 96 essays were used with a total of 20,891 words. The reference corpus is the Michigan Corpus of Upper-Level Student Papers (MICUSP). It is a corpus of A-graded papers by native and non-native English speakers at four levels of studies, and can be divided into four academic categories (Weigle & Friginal,
The present study uses papers by graduate students in that they are in line with what NETEM candidates are expected to produce in future studies.

Table 1. NETEM Writing and Disciplinary Writing

<table>
<thead>
<tr>
<th>Five domains</th>
<th>Number of papers</th>
<th>Tokens (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NETEM writing</td>
<td>96</td>
<td>20,891</td>
</tr>
<tr>
<td>Biological and Health Sciences</td>
<td>92</td>
<td>254,335</td>
</tr>
<tr>
<td>Humanities and Arts</td>
<td>78</td>
<td>374,569</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>79</td>
<td>248,383</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>150</td>
<td>494,020</td>
</tr>
</tbody>
</table>

3.2 Additive MDA and Dimensions of Linguistic Features

Biber’s (1988) dimensions have become stable references for additive MDAs across various domains (Nini, 2019; Berber Sardinha et al., 2019) in that this study described the overall patterns of register variation based on a wide range of language use situations. Additive MDAs allow researchers to focus on one dimension or several dimensions of interest in their studies. Another strength is that it makes research possible based on relatively small restricted corpus. This is the case in our study. The current study focuses on the first four dimensions because they have more explanatory power (Biber, 2019). The four dimensions are: Involved vs. Informational Production; Narrative vs. Non-Narrative Concerns; Explicit vs. Situation-Dependent Reference; Overt Expressions of Persuasion. In this study, 48 linguistic features were used in calculating mean scores of each dimension. Refer to Biber (1988, 2014) for details of the individual linguistic features. Dimension 1 and Dimension 3 include positive and negative features, which suggest they are distributed in a text in a complementary manner.

3.3 MAT and Data Analysis

All the fine-grained features on the four dimensions should be tagged and computed. Multidimensional Analysis Tagger (MAT) is an automatic tool developed by Nini (2019) that replicates the original Biber Tagger (not publicly available). It uses the Stanford Parser to parse lexical, syntactic and semantic information of the linguistic features in Biber (1988) and effectively replicates Biber’s algorithms and results. The reliability of MAT is an important driving factor for the application of Biber’s dimensions in additive MDAs (Berber Sardinha et al., 2019). This study uses MAT version 1.3.2 to map the five domains onto Biber’s 4 dimensions. We used MAT to complete 3 of the steps in Berber Sardinha et al. (2019): 1) tagging linguistic features in each text, 2) computing and standardizing their normed frequency counts, 3) computing the dimension scores for the five domains. SPSS was used to compare the differences in frequency counts for the linguistic features, and means on each dimension of the five domains. Despite the normal distribution of data in this study, Levene’s test of equality of variance was significant on all four dimensions. Therefore, Tamhane’s T2(M) post-hoc tests, which don’t assume equality of variance, were run to determine where significant differences were found.

4. Results

MAT classified NETEM writing as general narrative expositions while graduates’ papers as scientific expositions. This indicated NETEM writing used different language features from disciplinary papers. ANOVA results showed significant differences on all four dimensions: Dimension 1, $F(4, 220.24) = 43.223$, $p = .000$; Dimension 2, $F(4, 223.356) = 26.616$, $p = .000$; Dimension 3, $F(4, 219.012) = 18.350$, $p = .000$; Dimension 4, $F(4, 218.466) = 25.215$, $p = .000$. The means on four dimensions are displayed in Table 2 and visually shown in Figure 1, where the shared letters on each dimension represent mean scores that are not significantly different from each other. Post-hoc results with illustrative excerpts will be reported as follows.
Table 2. Descriptive Statistics (Means and SDs) across Dimensions and Domains

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Biological and Health Sciences</th>
<th>Humanities and Arts</th>
<th>Physical Sciences</th>
<th>Social Sciences</th>
<th>NETEM Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 1</td>
<td>-14.44 (5.54)</td>
<td>-11.16 (6.64)</td>
<td>-15.19 (4.57)</td>
<td>-13.06 (4.93)</td>
<td>-4.04 (7.23)</td>
</tr>
<tr>
<td>Dimension 2</td>
<td>-3.23 (1.54)</td>
<td>-1.86 (1.61)</td>
<td>-3.9 (1.36)</td>
<td>-2.51 (1.57)</td>
<td>-1.39 (3.36)</td>
</tr>
<tr>
<td>Dimension 3</td>
<td>7.58 (2.86)</td>
<td>5.42 (2.52)</td>
<td>6.56 (2.83)</td>
<td>7.64 (2.39)</td>
<td>4.69 (4.18)</td>
</tr>
<tr>
<td>Dimension 4</td>
<td>0.56 (3.81)</td>
<td>-1.06 (2.83)</td>
<td>-0.46 (2.90)</td>
<td>-0.42 (2.66)</td>
<td>4.71 (5.07)</td>
</tr>
</tbody>
</table>

Figure 1. Mean Scores on Four Dimensions across Domains

On Dimension 1, Table 2 shows the mean score of NETEM writing is -4.04, significantly higher than four disciplinary papers, indicating the involved and non-informational focus. This textual style can be illustrated through the reliance on present verbs, 1st person pronoun and possibility modals. Expert 1 illustrates the use of some of these features.

Excerpt 1: To be exact, mobile phones have occupied so much of our time that we even don’t have time to talk with people around us! In this sense, we can say that mobile phones have separated us from each other… (NETEM writing, Dimension 1 score = -9.29)

Contrary to these features, Excerpt 2 from Physical Sciences is characterized by negative features such as nouns, attributive adjectives, and prepositions. Meanwhile, word length as a negative feature suggests the use of advanced vocabulary in disciplinary writing.

Excerpt 2: If a simple transport model is sought, it is recommended that a spatially varying reaction rate be employed with a constant diffusion coefficient. Future work could incorporate hydrologic models to assess the impact of stormwater surges on contaminant transport. (CEE.G3.03.1 Dimension 1 score = -21.89)

On Dimension 2, the mean score of NETEM writing is -1.39. Figure 1 indicates no difference between NETEM writing and Humanities and Arts papers. They both used more narrative features than the other three disciplinary groups. Narrative features like 3rd person pronoun, present participial clauses are very common in NETEM writing (see Excerpt 3).

Excerpt 3: … he is convinced that it is his interest to explore the deep space that significantly contributes to his present achievement, making it easier for him to survive in this competitive commercial world. (NETEM writing, Dimension 2 score = 3.1)

On Dimension 3, NETEM scored 4.69, close to Humanities and Arts papers but significantly lower than the other three groups. This low score indicated more use of context-dependent features. Explicit references such as “nominalization and pied piping construction”
characteristic of informational writing are significantly less used in NETEM writing. Instead, it used more place adverbials and other adverbials as illustrated in Excerpt 4, while nominalization was found to be a salient feature in all disciplinary papers.

Excerpt 4: Many students are facing various choices ahead …studying abroad… (NETEM writing, Dimension 3 score = -0.26)

On Dimension 4, the mean of NETEM writing is 4.71. Figure 1 indicates NETEM writing is clearly distinct from all disciplinary writing. The higher the score is, the more persuasive the textual style is. NETEM writing used many linguistic features associated with persuasion such as infinitives and obligation modals as a way of suggesting what should be done to solve the problem mirrored in the prompt. Excerpt 6 is an example. Also, split auxiliaries are common in NETEM writing, usually in the form of formulaic expressions such as “as is vividly shown/illustrated in the picture” and “It is universally held that”, to name a few.

Excerpt 6: College students should not miss any opportunity to acquire knowledge to achieve their all-round development. (NETEM writing, Dimension 4 score = 5.83).

5. Discussion and Conclusion

Using corpus linguistic method MDA, the present study found that NETEM writing differed significantly from all four disciplinary domains on dimensions 1 and 4, and showed similarities with Humanities and Arts but differed from the other three domains on dimensions 2 and 3.

Regarding Dimension 1 and Dimension 4, NETEM writing was characterized by interpersonal interactions and persuasion, while disciplinary papers used more linguistic features associated with information density. These differences can be primarily explained by the differing purposes in writing tasks. Disciplinary tasks as a type of source-based writing focus on the presentation of propositional information while NETEM writing requires test takers to describe the cartoons vividly and argue convincingly for the viewpoints generated. Given that Dimension 1 is the most basic variation among registers in nearly all MDA studies (Biber, 2014) and therefore, an important marker for varying degrees of information density, a weak link exists between NETEM writing and real academic contexts. That is, it is impossible to predict how learners will perform linguistically on these two dimensions based on test scores (LaFlair & Staples, 2017). It is worth noting that the major syntactic functions of the linguistic features with negative weights in Dimension 1 are phrasal modifiers in noun phrases. The dense use of phrasal modifiers is the unique style of academic prose (Biber et al., 2011). Yet these features are disproportionately underrepresented in NETEM writing. Just as Kyle et al. (2021) pointed out “…if key linguistic features of a particular domain are not required for the successful completion of a task, support for the extrapolation inference is also weakened”, NETEM writing failed to elicit the language patterns required in graduate level studies.

The alignment of NETEM writing with Arts and Humanities papers on Dimension 2 and Dimension 3 is not surprising since this domain consists of many papers from philosophy, history and classical studies where narration and description are common. This finding is consistent with Weigle and Friginal (2015), Llosa et al. (2019), and Staples et al. (2018). However, similarities on these two dimensions do not support the extrapolation of test task to other disciplines because the sharp differences from disciplinary writing on other dimension(s) still make the authenticity of the task problematic (Staples et al., 2018).

In addition to the nature and requirements of NETEM writing, the stereotype of academic language as clausally complex is also at work, influencing the choices in NETEM writing of linguistic features different from disciplinary papers. For example, the bestseller among NETEM candidates published by a renowned press in China encourages the use of participial clauses and adverbial clauses as a way to score higher. Some education and training providers encourage test candidates to use more personal examples as an effective way to argue for viewpoints, thus making narrative features salient. All this is no different from placing an obstacle to learners’ academic language development.

The present study also affirms that independent writing alone is not representative of linguistic features that learners are expected to learn in the university setting. Although NETEM did undergo changes in test tasks, cartoons as writing prompts has been used since
the 1980s. It is time to re-examine and reform the test task. Admittedly, it is hard to solve the extrapolation problem once and for all. Yet test developers should give more attention to source-based writing skills, an important academic literacy in today’s higher education.

Acknowledgements

This study was supported by the 11th China Foreign Language Education Fund (ZGWYJYJJ11Z025) and Research on Philosophy and Social Sciences of Jiangsu Universities in 2022 (2022SJYB0101).

References


Enabling Visually Impaired People’s Chinese Literacy learning through Information Technology

Shelley Shwu-Ching YOUNGa & Jen-Li WANGb
aInstitute of Learning Science and Technology, National Tsing Hua University, Taiwan
bNational Center for High-Performance Computing
*scy@mx.nthu.edu.tw

Abstract: How to help the visually impaired people in Taiwan to be able to learn Chinese literacy online has been a big challenge. However, in order to overcome the obstacle, we intentionally designed and implemented a cloud-based Visually Impaired People (VIP) E-learning Platform that affords to empower their Chinese literacy learning via specific typing skills and relevant course design so that they would have the opportunity to become independent online navigators and learners as well. To facilitate and examine their performance in terms of typing accuracy and speed on the Chinese typing and information searching online, etc., a competition activity was organized. Both quantitative and qualitative data were collected and analyzed via a questionnaire administered to the 18 participants, divided into three groups based on their ages. The results indicate that overall the participants are highly satisfied with the VIP platform for its empowerment in Chinese literacy learning that affords their online independent learning opportunity and barrier-free communication skills with sighted people after using the VIP platform. Additionally, some suggestions for the VIP platform improvement were given by the participants to achieve a more user-friendly and better online learning environment in the future.

Keywords: Visually impaired people, e-learning, typing competition, Chinese literacy

1. Introduction

A computer-based multimedia learning environment is one of the most commonly used ways of learning nowadays. Multimedia learning emphasizes not only the use of visual aids but also auditory aids to improve learners’ learning and understanding (Mayer et al., 2002). Moreover, multimedia tools can be effective aids to the research consent process under some circumstances (Palmer et al., 2012).

However, for those who are visually impaired, the abundant visual aids provided by Multimedia are almost ineffective. The visual system is the main sensory source of human beings. Eight-seven percent of the total capacity of the brain process visual information (Hsieh, 2014). The limited channels for the visually impaired to understand and learn about the world is listening and tactility.

Providing equal learning opportunities for those who are visually impaired is getting more and more attention nowadays. Educational service organizations around the world have begun to provide online learning courses and digital learning resources for those visually impaired.

A university and an organization in Taiwan have jointly committed to the development of a cloud-based Visually Impaired People (VIP) E-learning Platform based on the scaffolding theory and Web Content Accessibility Guidelines (WCAG) 2.0 to help the visually impaired to learn online by using computer typing since 2016. It is hoped that through long-term research and promotion, the visually impaired could learn to benefit from this platform.
Therefore, given the length of the manuscript, this paper does not report the VIP system. Rather, this paper aimed to examine the potential of the visually impaired in Chinese typing and information searching through the competition results and questionnaire analysis. The following would be the research questions.

1. What was the participants’ engagement in terms of frequency of practicing, time spend, and frequency of typing the same materials?
2. How did visually impaired learners feel about using the VIP E-learning platform, in terms of advantages and the needed improvement?
3. What would be the performance of those participants in terms of typing accuracy and speed on the Chinese typing and information searching competitions?

2. Literature Review

2.1 Computer-based Multimedia Learning

Learners who receive information mainly via both auditory and visual channels simultaneously can absorb more sensory information, retain more memory, and increase their problem-solving ability (Mayer, 2005). Besides, learners can take advantage of five aids of multimedia to improve the effectiveness of learning, such as multimedia aids, contiguity aids, coherence aids, modality aids, redundancy aids in the computer-based learning environment (Mayer, 2002). However, apart from Modality aids, none of the other four seem to apply to the visually impaired since the other four rely on the visual aids. Moreover, the web pages nowadays consist of a large number of animations, special effects, or pictures that so-called multimedia, but no matter how attractive the visual aids are, the visually impaired still cannot benefit from them (Lin, 2003).

2.2 The Challenges for the Visually Impaired people in Typing Chinese

Contacting people and searching for information on the Internet using computer typing is undoubtedly a common activity and necessary skills for people living in the information era. For the visually impaired, computer typing is the main way for them to do text communication (Wang et al., 2014). However, the visually impaired encounter difficulties in typing in Chinese, such as the homophones with different meanings and the unavailability to check the correctness of words. Compared with English alphabetic systems, vocabulary is composed of letters and the pronunciation is basically as shown in spelling. For example, both “Male” and “Mail” are pronounced as “[meil],” and both “Whole” and “Hole” are pronounced as “[houi].” The spelling of the English words differentiates based on the context, but can be determined if the word is correctly used or not. On the other hand, Mandarin Chinese belongs to the logographic system, in which symbols represent the words themselves, and many words with homophones but different meanings are common. For example, both expectation (期待) and Umbilical cord (脐带) are pronounced as “qidai” in Mandarin Chinese; both stray and elk are pronounced as “miliu” “ku – ㄆㄢ” ; both stray (迷路) and elk (麋鹿) are pronounced as “miliu” “ㄌㄧˊ ㄌㄨˋ” in Mandarin Chinese but their meanings are different. This phenomenon will cause difficulties in computer text communication with a group of people and searching information on the Internet. In addition, the visually impaired cannot immediately distinguish the correctness of words by “seeing” when typing, so typing with typos is a common phenomenon (Wang, 2014; Weng, 2015).

3. Methodology

3.1 Introduction to the Competition

The competition for the visually impaired on Chinese typing and information searching was jointly organized by Taiwan Digital Talking Book Association (TDTB) and Graduate
Institute of Learning Sciences and Technologies, National Tsing Hua University, and sponsored by the foundation of Elite-well. The competition included two sections, Mandarin Chinese typing and information searching. In order to assure the competition quality, typing speed of at least 15 Chinese characters per minute is minimally required for the participants.

3.2 Rules for the competition

In the Chinese typing section, participants were given 10 minutes to type an article selected by the organizers with at least 1000 words presented via the NVDA report in the Mandarin Chinese twice, the higher record would be the final score. As long as the time is up, the VIP system will automatically stop and display the typing speed and accuracy. The difficulty of the selected articles encountered by the three groups of students in the competition increases according to their age.

In the information searching section, participants proceeded 20 minutes to search the correct answers to the designated questions designed by the organizers on the Internet and submitted the resource link. The difficulty of the questions encountered by the three groups of students in the competition increases according to their age. The more correct answers would win the competition.

The two sections of competition aimed to cultivate participants’ skills of computer operation and typing ability for further facilitate self-learning and satisfying career requirement.

3.3 Participants

The participants of this research are 18 people who participated in the typing competition for the blind on November 21, 2021. Their age ranges from 10 to 41 years old; education level ranges from elementary school to college or above.

3.4 Groups

1. Elementary school (under 12 years old): 2 participants with one male and one female
2. Junior high and senior school (13-18 years old): participants with 5 males and 3 females
3. Adult (above 18 years old): 8 participants with 4 males and 4 females

3.5 Four Phrases of this study

1. Preparation of competition materials, including selected articles for typing, and web searching questions
2. Questionnaire design
3. Competition: Chinese typing and information searching
4. Data collection and analysis

3.6 Research Methods

This research adopted mixed research methods, including both quantitative and qualitative approaches. The quantitative data were collected through the competition results of both Chinese typing and information searching and the questionnaire based on the 5-Point-Likert-Scale questions. The qualitative data were collected through the open-ended questions in the questionnaire, which was used to investigate the advantage and disadvantage of the VIP platform.

The questionnaire used in this research is designed with the Likert five-point scale to investigate the participants’ perception to the VIP platform, which was distributed after the end of the VIP typing competition on November 21, 2021.
3.7 Questionnaire analysis

Among the questionnaires returned, the valid questionnaire is 18. The following section would address research question 1: What is the participants' engagement in terms of frequency of practicing, time spend, and frequency of typing the same materials?

1) **Frequency of Practicing the Typing materials Per Week on the VIP platform**

   As shown in the Figure 1, among the 18 participants, 9 people use the VIP E-learning platform once or twice a week, accounting for 50% of the overall; 5 people watch the course 3-4 times a week, accounting for 28% of the overall; one person didn’t watch the course every week, accounting for 5% of the total; 3 people who have never watched the course, accounting for 17% of the total.

2) **Time on Each Chinese Typing Practice on the VIP Platform**

   As shown in the figure 2, among the 18 participants, 4 had never practiced, accounting for 22% of the total; 5 had practiced for 10-20 minutes each time, accounting for 28% of the total; 4 had practiced for 30-40 minutes each time, accounting for 22% of the total; 3 people practice 50 minutes to 1 hour each time, accounting for 17% of the total; 2 people practice more than 1 hour each time, accounting for 11% of the total.

3) **Frequency of Course Review on VIP Platform**

   As shown in Figure 3, among the 18 participants, 13 have never reviewed, accounting for 72% of the total; 2 have reviewed once a month, accounting for 11% of the total; 2 have reviewed twice a month, accounting for 11% of the total; 1 only watch the course two weeks before the game, accounting for 6% of the total. Two reasons for the high number of never reviewers are:
   1. They only use the platform as a place to practice typing.
   2. They don’t have time to review due to busy schoolwork or extracurricular activities.

The following sections would address the research question 2: How do visually impaired learners feel about using the VIP E-learning platform, in terms of advantages and the needed improvement.

4) **Satisfaction of Using VIP Platform**

   Eight satisfaction questions are provided in the form of the 5-point Likert scale. As shown in Table 1, apart from the discussion forum and online consultation, participants are highly satisfied with the VIP platform.

<table>
<thead>
<tr>
<th>Options</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convenience of accessibility</td>
<td>4.33</td>
<td>0.68</td>
</tr>
<tr>
<td>Learning method</td>
<td>4.11</td>
<td>1.07</td>
</tr>
<tr>
<td>Exercises after each course</td>
<td>4.07</td>
<td>0.86</td>
</tr>
<tr>
<td>Videos, text files, slides, and extended information</td>
<td>4</td>
<td>0.9</td>
</tr>
<tr>
<td>Discussion forum</td>
<td>3.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Online consultation</td>
<td>3.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Listening and typing exercises</td>
<td>4.6</td>
<td>0.48</td>
</tr>
<tr>
<td>Exercise of information searching</td>
<td>4.22</td>
<td>0.73</td>
</tr>
</tbody>
</table>

5) **Advantages of VIP Platform**

   The advantages of the E-Learning platform can be classified into two categories based on the results of the questionnaire, including the platform operability and the practicality of learning content.

   First, platform operability. Except from the elementary school group, 3 participants in junior & senior high school group and four participants in the adult group think the VIP platform provide a simple interface and easy to use, accounting for 39% of the total. For example, J2A: “This website is accessible, easy to operate, and easy to use.” C8C: “This website provides a simple interface and is easy to use.”

   Second, practicability of learning content. Forty-four percent of the participant are satisfied with the practicability of the learning content to empower them to type accurately,
accounting for 44% of the total. For example, E1A: “I become more familiar with the keyboard when operating a computer.” S1A: “I can practice typing speed and accuracy on the website.” J6A: “It can let the learners themselves know which words they type are incorrect, and allow those who are visually impaired know what the word means.” Three did not provide advice, accounting for 17% of the total.

6) Needed Improvements of the VIP Platform

The needed improvements can be categorized into three categories: information searching materials, interface of the VIP platform, and functions of the VIP platform.

First, information searching materials. Participants recommended to make the searching easier for Junior and Senior high school group.

Second, interface of the VIP platform. Participants recommended to add the automatic login function, improve the quality of online consultation, shorten the website URL of the VIP platform, make the whole webpage into the situation that the screen reader can read the words out.

Third, functions of the VIP platform. Participants recommended to classify the learning content according to difficulty, proofread the typos in the listening and typing articles, and update the article in a timely manner.

7) Competition Results

The competition was composed of two types of skills, including Chinese typing and information searching. Because typing with a high error rate will cause difficulty in communication or misunderstandings, the requirement for the error rate of the best typing record must be less than 6%.

First, Chinese typing competition. As shown in the Table 2, the best typing record in the Elementary school group is 430 words in 10 minutes, with an average of 38 words per minute and an error rate of 4%. The best typing record in the Junior & Senior high school group is 411 in 10 minutes, with an average of 36 words per minute and an error rate of 4%. The best typing record in the Adult group is 813 words in 10 minutes, with an average of 79 words per minute and error rate of 1%. Although JS1 in Junior & Senior high school group can type 698 words in 10 minutes with an average of 60 words per minute; and A1 in Adult group can type 1039 words in 10 minutes with an average of 86 words per minute, due to their 6% error rate, neither of them won.

Based on the Techficiency Quotient Certification (TQC) for Chinese typing, the Chinese typing speed can be divided into three levels according to the average typing speed per minute: Practical level: 15-29 characters per minute with the error rate less than 10%; Advanced level: 30- 79 characters per minute with the error rate less than 10%; Professional level: over 80 characters per minute with the error rate less than 10%. As we can see from Table 2, the best record for the Elementary group and the Junior & Senior high school group can reach to the advanced level. Besides, the best record in the adult can achieve to the professional level. Therefore, by diligent practicing on the VIP platform, the visually impaired can even perform better than those who are not visually impaired.

Second, information searching competition. The difficulty of the information searching differentiates depends on the age of each group. As shown in the Table 3, the best record in the Elementary school group can search 9 correct answers in 20 minutes. The best record in the Junior & Senior high school group can search 10 correct answers in 20 minutes. The best record in the adult group can search 6 correct answers in 20 minutes. Searching information through the Internet is a necessary skill nowadays, by practicing on the VIP platform, the visually impaired are empowered to have the ability to search information.

<table>
<thead>
<tr>
<th>Group/Numbers of participant</th>
<th>Winner</th>
<th>Total</th>
<th>Error rate</th>
<th>Average typing speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary school/2</td>
<td>E1</td>
<td>430</td>
<td>4%</td>
<td>38 words/min.</td>
</tr>
<tr>
<td>Junior &amp; Senior high school/8</td>
<td>JS2</td>
<td>411</td>
<td>4%</td>
<td>36 words/min.</td>
</tr>
<tr>
<td>Adult/8</td>
<td>A2</td>
<td>813</td>
<td>1%</td>
<td>79 words/min.</td>
</tr>
</tbody>
</table>

Table 3. The best record for each group in information searching
Group/Numbers of participant

<table>
<thead>
<tr>
<th>Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary school/2</td>
<td>9/15</td>
</tr>
<tr>
<td>Junior &amp; Senior high school/8</td>
<td>10/18</td>
</tr>
<tr>
<td>Adult/8</td>
<td>6/20</td>
</tr>
</tbody>
</table>

4. Conclusion

The results of this study indicate the effects of the modality aids proposed by Mayer. The online platform could benefit those VIP and they are highly satisfied with the VIP platform. Their performance on the typing speed and accuracy are outstanding and impressive. When the visually impaired is continuously practicing on the VIP platform, they can gradually reduce the typing error rate, which not only allows them to strength their information searching and communication skills, apply for jobs with this skill, but also reduces the inconvenience in communication with people in written language. Moreover, it is expected that the users on the VIP platform can take the advantage of this the practicing materials to achieve the self-directed learning on the Internet.

Acknowledgements

Thanks for the partial support from the Ministry of Science and Technology (MOST), R.O.C., under the grant codes: MOST 108-2511-H-007 -008 -MY3 and MOST 112-2410-H-007-0023, moreover, all of the participants involved in this study.

References


Tracking Knowledge for Learning Japanese as a 2nd Language

Tomoko OKIMOTO\textsuperscript{a}, Matthew JOHNSON\textsuperscript{b}, Huy NGUYEN\textsuperscript{c}, Steven MOORE\textsuperscript{d}, Michael EAGLE\textsuperscript{d} and John STAMPER\textsuperscript{c}

\textsuperscript{a}Tokyo University of Foreign Studies, Japan
\textsuperscript{b}Unaffiliated, Japan
\textsuperscript{c}Carnegie Mellon University, USA
\textsuperscript{d}George Mason University, USA
*tokimoto@tfus.ac.jp

Abstract: Most educational technologies for teaching language skills give students the same learning experience based on the assumption that every student learns at the same rate. In order to provide more individualized instruction we need to track student knowledge at a fine-grained level. This research explores how to add the ability to track knowledge in an existing educational technology system for Japanese second language learners. We explore several potential skill models based on the features available in the system and then apply a Bayesian knowledge tracing algorithm. We also make a large dataset for Japanese language learning available with no previous applied knowledge tracing.

Keywords: Knowledge Tracing, Skill modeling, 2nd Language Learning, Japanese

1. Introduction

This research is focused on improving the tracking of student second language learning knowledge from an existing question based training system for Japanese as a second language. Using established knowledge Tracing (KT) techniques, we were able to calculate the best fitting knowledge model and were able to show that by adding adaptive instruction, we could save learners from completing thousands of problems in the system that they have already mastered, allowing them to focus on skills that go unmastered.

The use of educational technology systems, such as online courseware and digital learning platforms, have been shown to be effective in helping learners acquire a second language (Webb & Doman, 2020). Research has shown that repeated practice is a key factor in successful second language learning, and many educational technology systems provide multiple opportunities for learners to practice and use the new words and skills they are learning (Nakata, 2017). Timely and corrective feedback is also an important feature of educational technology systems for second language learning, as it helps learners to identify and correct errors in real-time (Han, 2019). By leveraging adaptivity in these second language learning systems, the corrective feedback and multiple practice opportunities provided to students can be made more effective for their learning (Webb & Doman, 2020). One such way to enable this adaptivity is through KC Modeling, which not only enables adaptivity, but it also provides a way to accurately measure student knowledge. KCs are fine-grained representations of knowledge that include constraints, schemas, and production rules (Stamper & Koedinger, 2011). The KCs are typically mapped to every step of a problem, which is associated with an action the student needs to take in order to solve the problem. Each problem step provides an opportunity for students to demonstrate their mastery of the associated KCs. This mapping of KCs to the problem steps is known as a knowledge component model (KCM) and needs to be accurately designed to model student learning (Stamper and Koedinger, 2012). The use of knowledge tracing (KT) to model student learning through the progression of KCs has been proven effective in the domain of
language learning (Rizvi et al., 2022). Bayesian Knowledge Tracing (BKT) is a widely used KT algorithm, which uses parameters for learning rate, along with guess and slip parameters (Corbett & Anderson, 1994). This model is used in many intelligent tutoring systems and works for any domain.

2. System and Data Collection

The edtech system used in this study is designed to quiz Japanese second language learners and is focused on paired intransitive and transitive verbs. In the Japanese language, these verbs share a stem and have pairs with each other (e.g. あつめる (atsu-meru; gather something) and あつまる (atsu-maru, people gather)). The differences of those paired intransitive and transitive verbs are basically sentence structure and the particles used with them. Intransitive verbs have the subject while the transitive verbs do not require the subject. Even if Japanese language learners master those verbs at the beginner level, they are often common mistakes from intermediate or advanced level students as they progress (Okimoto, 2021). Since it is often difficult for Japanese language learners to master those paired verbs, Japanese language teachers often have difficulty building on these concepts without first addressing the difficulties students encounter (Nakaishi, 2020).

For the present study a total of 448 items were used in a 5-week course of online study, which intermediate level Japanese language learners joined. Participants answered a total of 50 questions per access, one access per day, five days a week. For the five item types, ten items were randomly selected. The average completion time for 50 items was approximately between 20 to 30 minutes. Items were given randomly and each item appeared 65 to 140 times in a 5-week period. Corrective feedback would appear after the student answered each item. There are two types of feedback, one for correct answers, the other for incorrect answers paired with explanations and correct answers.

The participants in this study were Japanese language learners who were enrolled in courses equivalent to the late intermediate to early advanced levels at an institution of higher education in Tokyo. There were 41 participants, 10 of whom were in the control group who took only the pre-test and post-test, and 31 in the experimental group who took the pre-test, 5-week online study, and post-test. Both the control group and the experimental group joined the Japanese language course at the institution. The breakdown of native languages was 25 Chinese native speakers, 5 Korean native speakers, 8 English native speakers, and 3 speakers of other Asian languages (Tagalog, Vietnamese, and Indonesian). The item dataset consists of 448 items, testing beginner and intermediate level paired intransitive and transitive verb usage of Japanese language learners. There are additional categorizations of the items: vocabulary difficult, JLPT-level, Intransitive or Transitive verb, Teramura verb (Teramura 1990), and motion verb. The vocabulary difficulty and JLPT-level are determined by the Matsushita database (Matsushita 2011). Teramura is either 0 or h, depending whether the verb is in the Teramura verb set. In the case that the target verb is in the Teramura dataset, it is labeled “h”, if it is not present in the dataset, it is labeled “0”. Lastly, motion verbs, 0 or m, describe whether the verb in the item is a motion based verb, a verb-type Japanese second language (JSL) learners have difficulty with.

3. Skill Models and Knowledge Tracing

Initial KCMs were created by using the features of the items in the system. We report the fit of five identified KCMs by Bayesian Information Criteria (BIC) and Item Blocked Root Mean Squared Error (IB RSME) in Table 1. BIC is a model comparison metric that is widely used to compare skill models in intelligent tutoring systems (Liu et al., 2016) and adds a penalty for the complexity of a model (Stamper et al., 2013). We use BKT to identify the number of opportunities that students need to reach mastery in each knowledge component, with the 22 KC model, which was the best by BIC. A student is considered as mastering a KC if their inferred mastery of this KC at the final attempt is at least 90% (Corbett & Anderson, 1994). Based on this notion, we found that all but 4 students mastered all 22 KCs where they were
given at least 30 or more opportunities on a given KC. Most of the KCs were mastered rather quickly suggesting that students are given more opportunities than they actually need and that adding adaptivity could improve learning time considerably. With the application of KC modeling and addition of Knowledge Tracing we show that we can more precisely track learning in our language question system.

Table 1. Proposed Skill Models based on the identified features evaluated by DataShop.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>KCs</th>
<th>BIC</th>
<th>IB RSME</th>
</tr>
</thead>
<tbody>
<tr>
<td>beginner-JLPTlevel-itemtype</td>
<td>22</td>
<td>27.963*</td>
<td>0.330442</td>
</tr>
<tr>
<td>item_type</td>
<td>5</td>
<td>27,975</td>
<td>0.32979*</td>
</tr>
<tr>
<td>beginner-JLPTlevel-itemtype-Teramura-motion</td>
<td>85</td>
<td>28,242</td>
<td>0.330839</td>
</tr>
<tr>
<td>beginner-JLPTlevel</td>
<td>5</td>
<td>28,456</td>
<td>0.332392</td>
</tr>
<tr>
<td>question_item_data</td>
<td>448</td>
<td>32,920</td>
<td>0.334551</td>
</tr>
</tbody>
</table>

Our next step is to use the KT to provide feedback to the students. A longer term goal is to add true adaptivity and intelligent tutoring capabilities to our language system in order to improve the learning efficiency of students using the system (saving thousands of problems). While this initial exploration was over several KC models with BKT for knowledge tracing, we plan to do more by exploring additional KT algorithms and run some pilot tests using the models and algorithms for adaptivity. Further, we have made the dataset available here in DataShop (Koedinger et al., 2013) [https://pslcdatashop.web.cmu.edu/Project?id=83](https://pslcdatashop.web.cmu.edu/Project?id=83), and hope others will also take advantage of this dataset that currently has many student attempts and no models or algorithms underlying the order of questions to the learners.

References


The Effect of Timing Differences in Online Corrective Feedback on Adult Verbal English Learners' Learning Engagement: A Micro-genetic Study

Wanying LIANG, Guang CHEN & Wei CHENG

Faculty of Psychology, Beijing Normal University, China
School of Educational Technology, Faculty of Education, Beijing Normal University, China
School of Educational Science and Technology, Nanjing University of Posts and Telecommunications, China

guang@bnu.edu.cn

Abstract: This study examines the impact of different timing direct corrective feedback strategies utilized by instructors in online verbal English class on the learning engagement of EFL adult learners. The study randomly divided 8 participants with the same English proficiency into two groups, one receiving immediate direct corrective feedback and the other receiving delayed direct corrective feedback. Applying a micro-genetic method, the study observed the online learning of the 8 participants and implemented two types of direct corrective feedback through online learning. The study found that both feedback methods could improve learners' learning engagement, but immediate direct corrective feedback had a more positive effect on learners' learning engagement, and could improve it in a shorter amount of time. Additionally, learners' learning engagement showed dynamic changes during the learning process due to other influencing factors.

Keywords: Micro-genetic method, immediate corrective feedback, delayed corrective feedback, learning engagement, online learning

1. Introduction

With the rise of globalization and international communication, more and more adults are learning English as a foreign language (EFL). EFL adult learners worldwide, particularly in the area of speaking, face various challenges. These challenges include difficulties in pronunciation, grammar, vocabulary, and cultural differences from their native language. Among this group of learners, feedback teaching strategies, including immediate and delayed corrective feedback, are widely used by teachers, and their effectiveness has been extensively studied. However, different scholars have reached varying conclusions regarding the impact of these two feedback strategies on students' learning. Some scholars believe that immediate corrective feedback is more helpful for learners' English language learning (Sakiroglu, 2020), while others argue that delayed corrective feedback is more effective (Nassaji & Kartchava, 2017). Similarly, some scholars have conflicting conclusions about the impact of different timing of corrective feedback strategies on learners' emotional states (Estaji & Farahanynia, 2019; Sarifah, 2019; Shabani & Safari, 2016a). These divergent research findings may stem from individual differences and the diversity of influencing factors, which further illustrates the complexity of the factors influencing learners' acceptance of feedback methods. In previous relevant research, the predominant traditional approach often centered on pretest-posttest methodologies, typically involving measurements before and after interventions. However, this approach didn't capture the changes in students' involvement during the learning process. Learning engagement is a complex process, and a new research method is needed to understand it better. The micro-
genetic method could be the solution to this gap in research. It's considered a good approach because it allows researchers to closely observe and analyze individual behaviors over time. It can help researchers gain a deeper understanding of the behavior and thought processes of EFL adult learners and provide more targeted learning support for them (Siegler & Crowley, 1991). On the other hand, learner engagement includes cognitive, emotional, and behavioral engagement (Fredricks et al., 2004). It can help observe changes in learner engagement under different feedback strategies, which can comprehensively understand the effectiveness of learners' learning, providing more guidance for teachers' use of teaching strategies. The fact that there hasn't been much research using the micro-genetic method to study how different feedback approaches affect learners' engagement highlights a gap in our knowledge. Clearly, studying the impact of various corrective feedback methods on how engaged learners are using the micro-genetic approach is very important and valuable for research.

2. Literature Review

2.1 The Impact of Types of Direct Corrective Feedback on EFL Learners

Direct corrective feedback refers to the teacher directly pointing out learners' errors and providing correct expressions. It can help learners correct pronunciation and grammatical errors and improve their verbal expression ability (Harmer, 2007). In addition, there are two main types of direct corrective feedback. The first is focused corrective feedback, which usually only corrects learners' errors in specific language targets or language skills to avoid overloading learners with too much information and unnecessary interference (Mackey & Goo, 2007; Nassaji & Swain, 2000; Sheen, 2007). The second is comprehensive corrective feedback, which refers to correcting students' language errors so that they can discover and correct errors on their own, thereby promoting their autonomous learning and language proficiency (Bitchener & Knoch, 2009; Ellis, 2009; Lyster & Ranta, 1997). Studies have shown that comprehensive corrective feedback has a positive effect on the learning performance of L2 learners (Cahyono, 2016; Van Beuningen et al., 2012) and can also significantly improve their learning interest and motivation, while focused corrective feedback has no significant effect (Falhasiri, 2021; Hartshorn & Evans, 2015). Therefore, this study will use direct comprehensive corrective feedback as the main feedback teaching strategy to observe the changes in learning engagement of EFL verbal English adult learners.

2.2 The Effects of Direct Corrective Feedback at Different Time Intervals on EFL Learners

The direct corrective feedback teaching strategy can be divided into two temporal dimensions: immediate direct corrective feedback and delayed direct corrective feedback. Immediate direct corrective feedback refers to correction given immediately after the learner produces an error, while the learner is still in the process of expressing themselves. Delayed direct corrective feedback, on the other hand, refers to correction given after the learner has completed the sentence or paragraph (Lyster & Ranta, 1997). Some studies have found that immediate direct corrective feedback can improve L2 learners' learning performance (Li et al., 2016; Tesnim, 2019). However, other studies have found that delayed corrective feedback can be more effective in improving L2 learners' fluency and accuracy of oral production (Hunter, 2012; Rahimi & Dastjerdi, 2012).

In addition to learning outcomes, the effects of direct corrective feedback at different time intervals also have different impacts on learners' emotional and affective states. Some scholars have found that learners receiving delayed feedback experience more negative emotions during the learning process, such as anxiety, frustration, and embarrassment (Quinn, 2014). However, other researchers have found that delayed feedback can help learners regulate emotions, improve learning outcomes, and have a positive impact on their...
self-worth and self-esteem (Estaji & Farahanynia, 2019). On the other hand, some studies have shown that immediate corrective feedback can help learners correct errors faster, and improve their affective state during the learning process, such as confidence and optimism (Yakisik, 2021). Interestingly, some scholars have drawn opposite conclusions. Immediate feedback may have negative effects on L2 learners’ confidence and produce negative emotion although it can help L2 learners improve their language proficiency (Shabani & Safari, 2016).

In summary, the effects of immediate and delayed corrective feedback on learners' emotions and affective states are complex and depend on various factors, such as individual learners, different subject areas, and learning environments.

2.3 The Application of Micro-genetic Method in Studying Learning Engagement

Micro-genetic method is a real-time analysis method for studying learning, which decomposes the learning process into small time units and analyzes the interaction between learners and the learning environment during these units. This method allows researchers to identify patterns and micro-changes in learner behavior, understand the basic mechanisms of learning, and study real-time dynamic processes in the learning process (Siegler & Crowley, 1991). Learning engagement refers to the degree to which learners actively participate in the learning process, including behavioral, emotional, and cognitive engagement (Fredricks et al., 2004). In recent years, micro-genetic methodology has been increasingly used to study the learning engagement of linguistic learners. Kaivanpanah and Miri (2018) used micro-genetic methodology to study changes in task-induced involvement of 24 Iranian EFL learners over an 18-month observation and recording period. The study found that cognitive engagement was influenced by multiple factors, such as task type, evaluation and teacher instructional style. In addition, emotional engagement predicted learners’ learning engagement. Jiang (2019) explored the changes in learning engagement and influencing factors of 12 EFL learners over a 10-week task interaction period using micro-genetic methodology. The results showed that learners' learning engagement fluctuated multiple times during the task process and was influenced by task type, language input and output, and other factors. Furthermore, learners generated many questions during the learning process, which promoted their learning engagement.

Therefore, learners are influenced by various factors during the learning process, leading to dynamic changes in learning engagement. Observing the trajectory of changes in learners’ engagement during the learning process can provide a comprehensive understanding of their learning situation, thereby facilitating the search for more effective teaching strategies.

3. Research Question

Based on the current research, most studies examining the effects of different timing for corrective feedback have primarily focused on the writing skills of EFL learners, while neglecting the exploration of oral proficiency development. Moreover, the application of micro-genetic methods to investigate changes in learner engagement among EFL learners remains largely unexplored. Therefore, this paper aims to use the micro-genetic method to observe changes in EFL adult learners’ engagement in online learning environments and analyze differences in their engagement trajectories after receiving direct corrective feedback at different timing. The goal is to gain a deeper understanding of the dynamic impact of these two types of feedback on learning engagement among adult EFL learners, and to compare the differences between them, so that more effective feedback approaches can be identified for instructional practices. To achieve this research objective, the following research questions are proposed:

1. Are there differences in the impact on EFL adult learners’ learning engagement between immediate direct corrective feedback and delayed direct corrective feedback teaching strategies?
2. In the teaching strategies of immediate direct corrective feedback and delayed direct corrective feedback, what are the specific trajectories of the impact on EFL adult learners' individual learning engagement?

4. Research Design

4.1 Participants

This study recruited eight participants whose English proficiency level was at the international A2 elementary level, as defined by the Council of Europe (Council of Europe, 2001). All participants had graduated from university several years ago and were engaged in different professions but had a personal need to improve their verbal English skills. Among the participants, there were two male and six female participants, ranging in age from 25 to 45 years old. The participants were randomly assigned to two experimental groups: one group received immediate direct corrective feedback instruction, and the other group received delayed direct corrective feedback instruction.

4.2 Instruments

4.2.1 Learning Engagement Scale

The Online Student Engagement Scale (OSE) (Cronbach's alpha = .95, p<.01) (Dixson, 2015) will be used in this study to measure students' learning engagement. This scale consists of four dimensions: skill engagement, emotion engagement, participation engagement, and performance engagement. Skill engagement refers to students' cognitive engagement in the learning process. Emotion engagement relates to students' emotional experiences and attitudes towards learning. Participation engagement pertains to the extent of students' involvement in online courses and tasks. Performance engagement reflects students' performance and grades in learning tasks. The scale comprises 19 questions and uses a 5-point Likert scale, ranging from "strongly disagree" to "strongly agree." As the experimental instruction in this study is one-on-one, the questions related to peer interaction in the participation dimension were modified to focus on interaction with the teacher.

4.2.2 Verbal English Task

The International English Language Testing System (IELTS) speaking exam is a comprehensive test that assesses test-takers' verbal proficiency. Studies have shown that IELTS speaking scores demonstrate high validity in predicting English language ability and actual performance (Li & Zhang, 2018). This study employed the first section of the Cambridge IELTS Speaking test for one-on-one online oral instruction. Each session lasted for 30 minutes, encompassing three distinct topics. Each topic consisted of four questions, with the primary instructional emphasis on grammar structures and oral expressions pertinent to the IELTS Speaking test.

4.2.3 Procedure

In this study, 8 participants were randomly divided into two groups: the immediate direct corrective feedback group and the delayed direct corrective feedback group. Four participants were assigned to each group. The study consisted of four one-on-one online oral English training sessions, which were conducted via Tencent Meeting. Prior to the experimental training, all participants received training in English sentence structure and related experimental instructions for the first part of the IELTS speaking test. During the experimental training, the examiner asked the participants English questions and they responded verbally. For the immediate direct corrective feedback group, the examiner immediately interrupted and corrected the participants when they made expression errors,
and then required them to repeat the corrected sentence before continuing with their response until the end of the question. For the delayed direct corrective feedback group, the examiner noted down the errors made by the participants during the oral response process and provided comprehensive corrective feedback after the completion of the entire topic. In addition, the examiner asked the participants to re-answer the questions. Written homework assigned by the teacher was completed before the start of the next class. At the end of each session, the participants completed a survey on their learning engagement and recorded a video of their responses to specific questions. The experimental training consisted of one session per week for four weeks.

### 4.3 Data Analysis

The online verbal English task in this study was conducted through Tencent Meeting software, allowing video recording. Participants’ learning engagement scores were collected through questionnaire responses, and further data collection on the reasons for changes in learning engagement was carried out using observational methods and self-reporting by the participants. Subsequently, the micro-genetic method was applied to encode and categorize the overall changes in learning engagement of the participants over time, facilitating a more comprehensive understanding of trends in engagement growth and decline. This method enabled an in-depth investigation into the impact of direct corrective feedback teaching strategies on participants’ learning engagement performance at different time intervals. The data analysis in this study was mainly conducted using the Excel 2021 version and the SPSS 27.0 version software. The learning engagement scores of the participants in the two experimental groups are presented and analyzed as below.

<table>
<thead>
<tr>
<th>Feedback Type</th>
<th>M</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediate Direct Correct</td>
<td>73.5</td>
<td>8.12</td>
</tr>
<tr>
<td>Delayed Direct Correct</td>
<td>70.25</td>
<td>6.23</td>
</tr>
</tbody>
</table>

Based on the average scores presented in Table 1, the immediate direct corrective feedback group had higher scores in learning engagement, while the delayed direct corrective feedback group had lower scores. Preliminary analysis suggests that immediate direct corrective feedback may have a more positive impact on participants’ learning engagement than delayed direct corrective feedback. Additionally, there were differences in the standard deviation and standard error between the two groups, indicating that the learning engagement scores in the immediate direct corrective feedback group were more stable, while those in the delayed direct corrective feedback group were more variable. Further analysis using the micro-genetic approach will be conducted to validate these findings.

### 5. Results and Discussion

#### 5.1 Comparison of Mean Score Changes in Learning Engagement for Two Direct Corrective Feedback Methods

As shown in Figure 1, the micro-analysis method was used to compare the average scores of learning engagement between the immediate and delayed direct corrective feedback groups in each class. The immediate direct corrective feedback group (3.25, 2.33, 2.67, 2.75) had slightly higher score changes throughout the process than the delayed direct corrective feedback group (1.25, -0.25, 4, 3), and the change trend was relatively stable. However, the score changes in the delayed direct corrective feedback group fluctuated in the first two classes, but increased significantly in the third and fourth classes, with average score changes higher than the immediate direct corrective feedback group throughout the
process. This indicates that the learning engagement of the immediate direct corrective feedback group is more stable than that of the delayed direct corrective feedback group in the learning process, while the delayed direct corrective feedback group's learning engagement improves significantly in the later stages of learning.

It can be seen that the immediate and delayed direct corrective feedback groups showed completely different trends in the average score changes of learning engagement in each class. The learning engagement of the immediate direct corrective feedback group steadily increased as the course progressed, while that of the delayed direct corrective feedback group fluctuated and was unstable. These data results may be related to the characteristics of the feedback modes of the immediate and delayed corrective feedback groups. The immediate direct corrective feedback group corrected students' errors in a timely manner and provided corresponding suggestions, which may help students make more correct responses in subsequent learning. However, the delayed direct corrective feedback group may not make students perceive their errors in a timely manner, leading to longer-lasting responses to errors and adverse effects on learning engagement.

**Figure 1. Average score changes in learning engagement between two different corrective feedback groups.**

5.2 Comparison of Rate Changes in Learning Engagement for Two Direct Corrective Feedback Methods

As shown in Figure 2, a microscopic analysis of the change rate reveals that the immediate direct corrective feedback group exhibited a higher change rate in the first two sessions, particularly with a change rate of 0.14 between the first and second session, while the delayed direct corrective feedback group showed a more stable change rate throughout the process with minimal fluctuations.

From this, it can be inferred that the immediate direct corrective feedback group demonstrated a gradually increasing trend in learning engagement scores throughout the course, with a higher change rate in the first two sessions. In contrast, the delayed direct corrective feedback group also showed improvement in learning engagement throughout the course, but with a more stable change rate. These findings suggest that immediate direct corrective feedback is more effective in producing positive effects on students' learning engagement in a shorter amount of time, while delayed direct corrective feedback may require a longer learning process to achieve significant improvement.
5.3 Variations in Individual Learning Engagement for Two Direct Corrective Feedback Methods

5.3.1 Individual Learning Engagement Changes in Immediate Direct Corrective Feedback

As illustrated in Figure 3, it demonstrates a general trend of improvement in the learning engagement of the four research participants over the course of the four sessions. Specifically, the first participant exhibited a score increase from 83 to 88, with relatively consistent change patterns. In the interview, the participant reported that the immediate direct corrective feedback facilitated a more profound understanding of incorrect language points in their verbal expression, thereby enhancing their motivation to learn, as evidenced by the reduced frequency of errors and a sense of achievement. Participant 2 demonstrated significant changes in scores between the first and second sessions, but subsequently displayed relatively stable changes, ultimately achieving a score of 68. During the interview, this participant indicated that the experiment had a positive impact on their English verbal skills. However, due to the establishment of a new company, they lacked the energy to devote to English learning and experienced interruptions from work during class. Participant 3’s score continued to rise steadily throughout the four sessions, increasing from 69 to 81. The participant reported that their work frequently involved the use of English and that the immediate direct corrective feedback helped them to deepen their understanding of language points, thus increasing their motivation to learn by reducing errors in verbal expression. Participant 4 demonstrated relatively stable score changes during the first three sessions but exhibited significant improvement during the fourth session, with a score increase from 77 to 84. During the interview, the participant stated that they initially felt pressure and lacked confidence when receiving instant direct corrective feedback, which resulted in hesitancy to express themselves. However, after adapting to this feedback method, their perspective changed, and they recognized it as an integral part of the learning process. They gained more confidence as they encountered and addressed challenges, ultimately becoming more willing to invest additional effort into learning.
5.3.2 Individual Learning Engagement Changes in Delayed Direct Corrective Feedback

As shown in Figure 4, the learning engagement scores of the fifth participant decreased from 74 to 65 between the first and fourth lessons. During the interview, Participant 5 reported having to work on a project that required a report to be submitted at the end of the experiment, resulting in overtime work and interruptions during the lessons due to work-related phone calls. This left the participant with insufficient energy to focus on learning. The participant had enough time for English language learning in the early stages of the experiment and perceived an improvement in language proficiency. However, the participant recognized the need to dedicate more time to learning. Participant 6 showed an increase in learning engagement scores for the first three lessons and a decrease to 63 in the fourth lesson. During the interview, Participant 6 reported feeling that their English verbal proficiency had improved with delayed and immediate corrective feedback. However, during the final week, the participant was occupied with end-of-quarter preparations for annual summaries and meetings, leaving little energy for language learning. Participant 7 showed a slight decrease in learning engagement scores for the fourth lesson, with a change from 64 in the first lesson to 70 in the fourth lesson. The participant reported being less confident during the first lesson, which led to reluctance to speak, but as the course progressed, the participant gained more confidence and willingness to learn. However, the participant's workload increased during the final week, leading to a reduction in English language learning time. Participant 8 showed a continuous increase in learning engagement scores across all four lessons, with a change from 69 in the first lesson to 79 in the fourth lesson. During the first lesson, Participant 8 expressed doubts about the effectiveness of delayed and immediate corrective feedback and believed that learning English grammar systematically would be more meaningful. However, in the subsequent lessons, the participant found that their English verbal expression and language proficiency had gradually improved, resulting in greater learning motivation and willingness to dedicate more time to learning.

![Figure 4 Individual changes in learning engagement for delayed direct corrective feedback group.](image)

5.3.3 Comparison of Changes in Learning Engagement of Individuals for Two Corrective Feedback Methods

Overall, there was a positive trend in the learning engagement of all participants, indicating that both types of direct corrective feedback had a positive impact on their learning engagement. However, immediate direct corrective feedback had a more significant positive effect. Notably, the significant decrease in learning engagement in the final week observed in participants 5 and 6 was attributed to work-related factors, whereas the continuous increase in learning engagement in participants 1, 3, and 4, and 8 was influenced by their learning motivation and sense of achievement. These findings suggest that learners' learning engagement can be influenced by various factors, such as learning motivation, work, and cognitive awareness, etc., resulting in fluctuating changes in their learning engagement.
6. Conclusion

This study compared the effects of immediate corrective feedback and delayed corrective feedback on learners' learning engagement and found that both types of feedback significantly increased learning engagement. However, the immediate corrective feedback group had higher scores in learning engagement than the delayed corrective feedback group. The changes in learning engagement also differed between the two groups, with the immediate corrective feedback group showing more significant changes than the delayed corrective feedback group, which is consistent with previous research findings. Masià and Ramon's study (2013) found that immediate feedback can improve learning engagement and learning outcomes. Meanwhile, Winstone et al. (2017) emphasized the importance of timely feedback for student learning. The results of this study suggest that immediate corrective feedback can improve the average score of learning engagement and can change learning engagement in a relatively short period. However, delayed corrective feedback is also effective in improving learning engagement, especially for students who need more time to understand and digest information. Additionally, in the process of learners' learning, learning engagement can be influenced by other factors, especially for working EFL adult learners who are affected by work-related factors. Learners should reflect on their learning situation in a timely manner and plan for frequent learning in their free time to improve their learning engagement and learning outcomes. Teachers should also consider other influencing factors in the teaching process and provide timely feedback to students so that they can understand their learning situation and make corresponding adjustments.

Future research can increase the sample size to further verify the generalizability of the conclusions. Additionally, it is possible to further explore the micro-level effects of different types of feedback on various sub-dimensions of learners' learning engagement. On the other hand, it is also possible to investigate how different types of feedback affect learners' learning engagement in different scenarios, such as small or large classes. Furthermore, it is possible to further investigate the effects of different types of feedback on the learning outcomes and learning engagement of learners with different personalities. These studies will help to better understand the role of feedback in student learning and provide better support for improving student learning outcomes.

References


Using Self-Regulated Digital Storytelling in Primary Students’ English Learning: An Exploratory Factor Analysis

Yunsi Tina MAa,b*, Siu-Cheung KONGa & Daner SUNa
aDepartment of Mathematics and Information Technology,
bCentre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong
*yma@eduhk.hk

Abstract: Successful learners are cultivated by good learning habits and essential skills. Among them, self-regulated learning (SRL) is proven to be an effective learning strategy and pedagogical approach to develop independent learning, enhance learning outcomes, improve learning motivation, and finally achieve lifelong success. In language education, Digital Storytelling (DST) has been an emerging strategy for teachers to guide their students to learn a language by expressing ideas and meanings. However, little research has been done to explore the effectiveness of such integration on students’ language learning. The present study designed and developed a self-regulated digital storytelling (SRDST) curriculum for teachers to teach and students to learn English writing. About 110 primary grade 4 students of 9-10 years old participated in this study. A SRDST scale with 13 items in forethought, performance, and reflection phases was developed and validated. The results of exploratory factor analysis (EFA) from the primary students (N=70) on a pre-/post-test design yielded three factors that were similar to the three phases of SRL, namely, forethought, performance and reflection stages. The paired-samples t-test results indicated a significant increase in forethought, suggesting that the students realized certain preparation needs to be done before their study, and they also preferred autonomy on managing their own learning. Further, the data demonstrated a significant result in the performance phase where the students can seek for help when encountering difficulties, and determine to keep up with the overall learning progress and complete the tasks on time. The results also revealed that the students tended to reflect on whether there were better ways to improve their final projects. Finally, the research found that young learners preferred to set learning goals during the performance stage which is different from adult learners.

Keywords: Self-regulated learning (SRL), digital storytelling, English learning, instruments, exploratory factor analysis (EFA)

1. Introduction

Self-regulated learning (SRL) is not only a learning strategy to support student learning but also a pedagogy to facilitate teacher teaching. SRL can be used as a stand-alone pedagogy or combined with other pedagogical approaches to facilitate student-centred learning (Barrett, 2006). While many researchers investigated the effectiveness of SRL in higher education and adult education both offline and online in the past years (Carter Jr et al., 2020; Jansen, 2019; Vanslambrouck et al., 2019;), more studies have been probing the impact of SRL in online or blended learning mode after the outbreak of COVID-19 pandemic when researchers and educators uphold the importance of SRL abilities. Previous studies argued that SRL can improve young learners’ learning outcomes, learning strategies and motivation (McClelland & Cameron, 2011; Hung et al., 2012; Järvelä et al., 2012). Dignath et al. (2008) pointed out that primary school students can also acquire SRL and benefit from it. Researchers believe that SRL can foster sustainable lifelong learning skills as SRL can cultivate learners’ generic abilities, such as problem-solving skills, digital competencies and learning autonomy (Anthonysamy et al., 2020; Lüftenegger et al., 2012). Recently, the application of SRL in
In the last four decades, educators and researchers probed the impact of digital storytelling as a pedagogy in various subject disciplines across primary, secondary, and higher education levels (Wu & Chen, 2020). Digital Storytelling (DST) is a technology-support strategy for teachers to guide their students to learn a foreign or second language by expressing ideas and meaning through integrating the use of audio, texts, and visual artefacts to develop language ability and communication skills (Wang & Zhan, 2010; Torres et al., 2012). Besides, it is also an innovative narrative tool for promoting digital competencies and relevant skills (Del-Moral-Pérez et al., 2019). Moreover, DST can also be a learning strategy for improving learners’ academic achievement, problem-solving skills, learning motivation and collaboration skills (Hung et al., 2012; Niemi & Multisilta, 2016). In the field of language learning, DST has proved to be one of the most popular and effective strategies for improving students English writing (Yang & Wu, 2012; Burke & Kafai, 2010). However, the research on innovative pedagogies or tools for facilitating primary students’ DST is limited. Furthermore, aiming to equip students with digital competence in language learning, the current study aims to explore whether the SRDST enabled by technologies is effective in grooming young learners’ digital competency development. Thus, for addressing the above issues, the integration of programming tools in DST, for instance, Scratch, will have great potential for engaging students in SRL learning environment.

2. Literature Review

2.1 Self-regulated Learning

Zimmerman (2002) concluded that self-regulation is neither a mental ability nor an academic performance skill. Instead, it is a process that learners self-direct themselves to transform their mental capabilities into academic skills and achievements. Self-regulated learners can apply multiple strategies including cognitive strategies (e.g. organization and elaboration), metacognitive strategies (e.g. planning and monitoring), resource management strategies (e.g. time management), and collaborative learning to improve their learning (Bai et al., 2021). Self-regulated learning can also develop positive thinking by guiding students to have an optimistic future planning which can guide them to overcome obstacles and succeed academically (Zimmerman, 2002). Therefore, SRL is essential for grooming lifelong learning skills in a digital world (Anthonysamy et al., 2020). Further, Loyens et al. (2008) found that the “self” aspect is important in the developmental processes of SRL. Therefore, emphasizing student-centred learning also aligns with the gist of SRL.

Riding on Bandura’s social cognitive theory (1986) interplaying person, behavior and environment, researchers developed and entailed various notions into SRL. Pintrich (2004) added metacognition and social context to the SRL framework and expanded the regulating behaviors to the forethought phase. Forethought includes task analysis and self-motivation belief, performance includes self-control and self-observation, and self-reflection includes self-judgement and self-reaction (Zimmerman, 2008). Meanwhile, Schunk (1990) pointed out that students shall be taught to set realistic upper and lower goals to enter learning activities for goal attainment with self-efficacy. In the SRDST process, the students have little idea about the activities in the forethought period, therefore, the current research attempted to put goal setting under the early performance stage so that the students can set and adjust the upper goals or main goals, and lower goals or sub-goals throughout the performance stages. There have been various SRL models working on Zimmerman’s three phases, namely, forethought, performance, and self-reflection (Panadero, 2017). In our study, we adopted an adapted Zimmerman’s phases (Zimmerman & Moylan, 2009) entailing different self-regulated behaviours aiming to propose one framework that fits the primary school level. In the framework for the primary level, the forethought phase entails environment, learner autonomy and strategic planning, performance phase entails goal-setting, time management, self-efficacy, help-seeking and progress monitoring, and the reflection phase entails self-reflection
with or without peer assessment, learner engagement, improvement and well-being (Figure 1).

Figure 1. SRL framework proposed by the current study for primary levels based on Zimmerman and Moylan’s framework (2009)

2.2 Self-regulated Digital Storytelling

DST facilitates a series of student-centred learning strategies, namely, student engagement, reflection for deep learning, project-based learning and effective integration of technology with learning and teaching by using technologies in a meaningful way in classroom settings (Barrett, 2006; Stale & Freeman, 2017). Wu and Chen (2020)’s systematic review on DST for educational purposes revealed that DST had been applied in various levels of educations ranging from primary, secondary to higher education levels as either a stand-alone pedagogy or in combination with other pedagogies, usually in culture, gender, language, social psychology, and social studies and 21st century skills among which language learning is the main subject where DST is employed. Lim et al. (2022) observed that little research offered detailed definitions of related research topics, concepts and methods of digital storytelling for language learning. Therefore, the current study proposes an SRDST framework and model to guide the research activities aiming to fill in the research gaps and probe the impacts of SRDST on EFL/ESL primary students’ learning in English writing (Ma et al., 2022). In gist, SRDST is a digital storytelling approach steered by SRL for enhancing learning outcomes and improving learning motivation.

A few studies found that digital storytelling on Scratch had a positive impact on enhancing learner’s motivation and academic achievement in English learning. Digital storytelling on Scratch offers middle school learners the opportunity to better understand the process of developing a story from an idea to a digital story (Burke & Kafai, 2010; Parsazadeh et al., 2021). Smith and Burrow (2016) suggested ways to integrate DST into the classroom to support reasoning, creative thinking and problem-solving. By partnering with teachers from language arts, technology and math subjects, Wolz et al. (2011)’s study revealed that teachers can infuse computational thinking in language learning in technology-enhanced classroom settings so as to enrich the language arts curriculum. Burke and Kafai (2012) affirmed that there is a great deal of learning potential in this intersection of the formal writing practices taught during the school-day and the “informal” activities of digital creation.

2.3 Background of the SRDST Scale

Previous research attempted to develop and validate SRL scale embedded with other learning strategies in different contexts while mainly in the realm of higher education (Roth et al., 2016). Barnard et al. (2009) developed a six-dimension questionnaire to measure SRL in online or hybrid learning modes demonstrating environment structuring, goal setting, time management, help-seeking, task strategies, and self-evaluation. Although this six-dimension
questionnaire is slightly different from Zimmerman’s three phases by subdividing forethought and performance phases, the current research adopted and adapted both frameworks as they match the results and findings from the previous curriculum design using a design-based research approach (Ma et al., 2023; Zimmerman, 2002).

Pintrich (2004) and Magno (2011) developed a framework to classify the phases and relevant scales based on the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991) in higher education. Retnawati (2016) developed an SRL instrument consisting of a Likert scale and multiple-choice questions for college students who majored in mathematics education. Based on their self-efficacy for self-regulated learning scale for elementary to high school learners, Usher and Pajares (2008) found that elementary students attained higher self-efficacy for self-regulated learning than those in higher levels. Roth and her colleagues (2016) also reported that since 1980s there had been an emerging development and deployment of subject-specific questionnaires over the past decades. Besides, they also claimed that MSLQ seemed to be the most verified instrument in SRL research, however, the MSLQ may not be able to assess post-performance behaviours such as self-reflection and regulation. Riding on the literature review and findings, the present research developed a SRDST scale to evaluate students’ performance and improvement in learning English writing guided by SRDST strategies.

2.4 Research Purposes and Research Questions

The current research, on the one hand, aims to develop and evaluation an innovative pedagogy, i.e. SRDST, for facilitating primary students’ DST in English learning, on the other hand, it aims to equip students with digital competence in language learning by deploying the SRDST strategy. To achieve these research objectives, the current study attempts to address the following research questions (RQs):

(1) What are the factors related to SRDST as perceived by the primary 4 students? Are they similar to the factors identified in the literature?
(2) Do the students perceive an improvement in the factors after attending the SRDST for English learning curriculum?

3. Method

3.1 Participants

About 110 EFL/ ESL Chinese students of 9 – 10 years old from a Hong Kong local primary school participated in this study. The Human Research Ethics Committee of The Education University of Hong Kong considered and approved this study by reviewing the research methods, procedures and instruments prior to the start of the research activities. The participating students provided approved written consent from their parents. The students were divided into one experimental group and one control group. The experimental group were instructed based on SRDST approach while the control group were instructed based on paper-formed worksheets and presentations slides.

3.2 The Self-Regulated Digital Storytelling Curriculum

Figure 2 demonstrates the rationales of the lesson design and the cycling process of the SRDST approach. The SRDST emphasizes the three phases of SRL: forethought, performance I and II and reflection to guide students to acquire the vocabulary, grammar and writing skills. Guided by SRDST approach, Table 1 presents a lesson plan for a primary grade 4 English Language unit “We Love Festivals” where SRDST strategies were embedded throughout the curriculum and teaching and learning resources. The experiment was conducted in four weeks in the first semester of 2022/23 academic year. The development of
the curriculum followed a design-based research approach and was guided by the SRDST innovative pedagogy proposed by this research (Ma et al., 2023).

Figure 2. Using SRDST approach to learn and teach the P4 English Language unit “We Love Festivals”

Table 1. A lesson plan of the P4 English Language unit “We Love Festivals”

<table>
<thead>
<tr>
<th>Lesson</th>
<th>Duration*</th>
<th>Activities</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>30 mins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesson 0</td>
<td>25 mins</td>
<td>Vocabulary</td>
<td>Flipped Learning</td>
</tr>
<tr>
<td>Lesson 1</td>
<td>45 mins</td>
<td>Grammar 1</td>
<td>English Class</td>
</tr>
<tr>
<td>Homework</td>
<td>Lesson 1 Take-home Assignment</td>
<td></td>
<td>Homework</td>
</tr>
<tr>
<td>Lesson 2</td>
<td>45 mins</td>
<td>Grammar 2</td>
<td>English Class</td>
</tr>
<tr>
<td>Homework</td>
<td>Lesson 2 Take-home Assignment</td>
<td></td>
<td>Homework</td>
</tr>
<tr>
<td>Lesson 3</td>
<td>45 mins</td>
<td>Writing: To create your own story</td>
<td>English Class</td>
</tr>
<tr>
<td>Homework</td>
<td>Lesson 3 Take-home Assignment</td>
<td></td>
<td>Homework</td>
</tr>
<tr>
<td>Lesson 4</td>
<td>45 mins</td>
<td>Writing: To create your own story</td>
<td>English Class</td>
</tr>
<tr>
<td>Homework</td>
<td>Lesson 4 Take-home Assignment</td>
<td></td>
<td>Homework</td>
</tr>
<tr>
<td>Lesson 5</td>
<td>45 mins</td>
<td>To create the story on Scratch</td>
<td>Computer Class</td>
</tr>
<tr>
<td>Lesson 6</td>
<td>45 mins</td>
<td>To create the story on Scratch</td>
<td>Computer Class</td>
</tr>
<tr>
<td>Lesson 7</td>
<td>45 mins</td>
<td>Presentation</td>
<td>English Class</td>
</tr>
<tr>
<td>Post-test</td>
<td>30 mins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>340 mins</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3 Instruments and Procedure

Riding on the literature review, the current study developed and employed an SRDST scale with 13 items of forethought, performance and reflection to evaluate students’ performance and improvement in learning English writing guided by SRDST strategies. In this study, the samples would be considered as two equivalent groups because (1) the distribution of gender, age, and class size were similar, (2) the teachers, textbooks, and time spent on the English classes for both groups were all the same, and (3) other important confounding variables, including students’ DST experience, years of learning English, coding experience on Scratch were not significantly different based on the results of a chi-square test for homogeneity.

4. Results

4.1 Descriptive Statistics

Among the 110 participants, 70 completed both the pre- and post-tests and regarded as the data sample. The sample students consisted of 38 females and 32 males from primary grade 4 whose ages ranged from 9-10 years old (M = 9.21; SD = .41). The English language learning
experience of the students ranged from 6 to 7 years. About half of the students have some coding experience on Scratch and the other half has no coding experience at all (M = .51; SD = .50). It is assumed that there is no significant difference between the two groups of students. The descriptive statistics of the two groups are presented in Table 2. The chi-square test of independence showed that there was no significant difference between the two groups in terms of gender, \( X^2 (1, N = 70) = 1.490, p = .22 \), age, \( X^2 (1, N = 70) = 1.775, p = .18 \), and experience of coding, \( X^2 (1, N = 70) = .525, p = .47 \). Besides, the English teachers' year of experience were also similar. Therefore, any differences between the experimental group and the control group could be regarded as independent variables in the current research.

<table>
<thead>
<tr>
<th>Table 2. Descriptive Statistics of the Two Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = 70</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>All Students (N = 70)</td>
</tr>
<tr>
<td>Control (n = 34)</td>
</tr>
<tr>
<td>Experimental (n = 36)</td>
</tr>
</tbody>
</table>

Note. N = Gender (1 = male; 0 = female); Coding Experience (1 = Yes; 0 = No)

4.2 Reliability Test and Factor Analysis

A Cronbach's alpha reliability test was conducted to measure the reliability of the SRDST scale with the pre-test data. According to the test results presented in Table 3, the reliability of the questionnaire was robust (Cronbach's \( \alpha = .858 \)). Its 13-item scale examined three phases of SRDST in an English writing class: a) forethought (three items); b) performance (six items); and c) reflection (four items). It was measured using a 5-point Likert scale, ranging from 1, 'Strongly disagree' to 5, 'Strongly agree'. In this study, the reliability of the scale was satisfactory (forethought: Cronbach's \( \alpha = .610 \); performance: Cronbach's \( \alpha = .822 \); reflection: Cronbach's \( \alpha = .822 \)) as presented in Table 4.

To explore the factor structure of the SRDST instrument, a factor analysis was conducted on the pre-test data collected from the students (N = 70) of the 13 items using SPSS, with Principal Component Analysis and Varimax with Kaiser Normalization as the extraction and rotation methods, respectively, which yielded three factors that were similar to the three phases of self-regulated learning, namely, forethought, performance and reflection stages, as shown in Table 4. The sample size met the minimum requirement of delivering an EFA (Bryant & Yarnold, 1995; Winter et al., 2009). The Kaiser–Meyer–Olkin value was 0.83, and the chi-square value for Bartlett's test of sphericity was 375.11 (df = 78, p < .001), indicating that the three factors had good explanatory power. The total variance accounted for by these three factors was 62.79%. The factor loadings of the items are presented in Table 4.

<table>
<thead>
<tr>
<th>Table 3. Reliability and factor loading of the SRDST scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-regulated Digital Storytelling Scale (N = 70)</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>SRDST Scale</td>
</tr>
<tr>
<td>Forethought</td>
</tr>
<tr>
<td>1. I know where and when I can learn English most efficiently.</td>
</tr>
<tr>
<td>2. I consult or discuss with someone when I need help in my learning.</td>
</tr>
<tr>
<td>3. I have the autonomy to achieve my learning goals in my own way.</td>
</tr>
<tr>
<td>Performance</td>
</tr>
<tr>
<td>4. I proactively seek for relevant information to understand the self-regulated digital storytelling activities on Scratch if I have difficulties.</td>
</tr>
<tr>
<td>5. I have set sub-goals (e.g. read the digital stories on Scratch on time and complete the worksheets) leading to the main learning goals.</td>
</tr>
<tr>
<td>7. I have enough time to learn English.</td>
</tr>
<tr>
<td>8. I make sure I can keep up with the overall progress of self-regulated digital stories on Scratch for learning English (e.g. read the digital stories on Scratch on time and complete the worksheets.</td>
</tr>
</tbody>
</table>
9. I can schedule a suitable time to read the digital stories on Scratch for learning English. 
10. I receive guidance on how to achieve the goals in stages. 
11. I enjoy the parts of Scratch digital stories and worksheets that lead me to reflect on my English learning. 
12. I will reflect on what I have learned after I finish the self-regulated digital storytelling activities. 
13. After I finish the self-regulated digital storytelling activities, I will reflect on whether there are better ways of creating my digital stories.

Note. A 5-point Likert scale is used, where 5 = “strongly agree”, and 1 = “strongly disagree”. *p ≤ .05; **p ≤ .01; ***p ≤ .001

4.3 Paired-samples T-test

Among the participants, 52 of them completed the whole experiment process by attending all the learning activities and finished the pre-/post-test surveys. The others who were absent from one or some classes were excluded from the paired-samples T-test analysis. The means, standard deviations and t-values are presented in Table 4.

Table 4. Means, standard deviations and t-values of the Paired-samples T-Test

<table>
<thead>
<tr>
<th>Self-regulated Digital Storytelling Scale (N = 52)</th>
<th>Pre-test Mean</th>
<th>SD</th>
<th>Post-test Mean</th>
<th>SD</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forethought</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. I know where and when I can learn English most efficiently.</td>
<td>2.670</td>
<td>1.642</td>
<td>3.620</td>
<td>1.207</td>
<td>-3.424**</td>
</tr>
<tr>
<td>2. I consult or discuss with someone when I need help in my learning.</td>
<td>3.080</td>
<td>1.296</td>
<td>3.830</td>
<td>1.115</td>
<td>-2.819**</td>
</tr>
<tr>
<td>3. I have the autonomy to achieve my learning goals in my own way.</td>
<td>2.920</td>
<td>1.326</td>
<td>3.710</td>
<td>1.126</td>
<td>-3.308**</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I proactively seek for relevant information to understand the self-regulated digital storytelling activities on Scratch if I have difficulties.</td>
<td>3.060</td>
<td>1.211</td>
<td>3.790</td>
<td>1.304</td>
<td>-2.919**</td>
</tr>
<tr>
<td>5. I am clear about the main learning goals of self-regulated digital storytelling on Scratch.</td>
<td>3.330</td>
<td>1.396</td>
<td>3.460</td>
<td>1.335</td>
<td>-0.555</td>
</tr>
<tr>
<td>6. I have set sub-goals (e.g. read the digital stories on Scratch on time and complete the worksheets) leading to the main learning goals.</td>
<td>3.170</td>
<td>1.410</td>
<td>3.270</td>
<td>1.315</td>
<td>-0.467</td>
</tr>
<tr>
<td>7. I have enough time to learn English.</td>
<td>3.060</td>
<td>1.243</td>
<td>3.440</td>
<td>1.145</td>
<td>-1.853</td>
</tr>
<tr>
<td>8. I make sure I can keep up with the overall progress of self-regulated digital stories on Scratch for learning English (e.g. read the digital stories on Scratch on time and complete the worksheets).</td>
<td>2.960</td>
<td>1.533</td>
<td>3.420</td>
<td>1.194</td>
<td>-2.299*</td>
</tr>
<tr>
<td>9. I can schedule a suitable time to read the digital stories on Scratch for learning English.</td>
<td>3.100</td>
<td>1.347</td>
<td>3.190</td>
<td>1.253</td>
<td>-0.437</td>
</tr>
<tr>
<td>Reflection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I receive guidance on how to achieve the goals in stages.</td>
<td>3.130</td>
<td>1.329</td>
<td>3.290</td>
<td>1.242</td>
<td>-0.704</td>
</tr>
<tr>
<td>11. I enjoy the parts of Scratch digital stories and worksheets that lead me to reflect on my English learning.</td>
<td>2.790</td>
<td>1.160</td>
<td>3.120</td>
<td>1.215</td>
<td>-1.761</td>
</tr>
<tr>
<td>12. I will reflect on what I have learned after I finish the self-regulated digital storytelling activities.</td>
<td>3.080</td>
<td>1.296</td>
<td>3.270</td>
<td>1.254</td>
<td>-0.919</td>
</tr>
<tr>
<td>13. After I finish the self-regulated digital storytelling activities, I will reflect on whether there are better ways of creating my digital stories.</td>
<td>3.100</td>
<td>1.376</td>
<td>3.630</td>
<td>1.189</td>
<td>-2.442*</td>
</tr>
</tbody>
</table>

Note. A 5-point Likert scale is used, where 5 = “strongly agree”, and 1 = “strongly disagree”. *p ≤ .05; **p ≤ .01; ***p ≤ .001

5. Discussions

According to the factor analysis results, we calculated the pre-test and post-test means of the three factors and carried out a paired-samples T-test to examine whether students have significant improvements after the completion of the course. The t-test results of the three factors, together with those of individual items, were as shown in Table 4. On the one hand, the results showed a significant increase in forethought, suggesting that the students realized certain preparation needs to be done before their study. With learning activities facilitated by technologies and internet, the students learned to find a suitable place and a good time to...
attend the activities at home. SRDST approach encouraged students to seek for help when
they were in need. They also appealed for autonomy in managing their own learning and
achieving their goals in their own ways. SRL has a positive impact on learners' autonomy and
under some conditions can support learners' autonomy (Papamitsiou & Economides, 2019;
Yamauchi & Tanaka, 1998). Sierens et al. (2009) argued “when teachers want their students
to evaluate themselves, to plan their study activities, and to think about themselves as learners,
the teachers are encouraged to provide help, instructions, and expectations in an autonomy-
supportive way.” On the other hand, the data demonstrated a significant result in the
performance phase where the students can seek relevant information when they encounter
difficulties and determine to keep up with the overall progress for learning English and
complete the tasks on time. The abilities and habits of keeping up with the learning progress
are very critical to ensure learning effectiveness (Hong et al, 2021).

Although Zimmerman’s framework (2002) puts goal setting in the forethought phase, in
the current study, it is found that the young learners set and adjusted sub-goals during the
performance phase to guide them to achieve the main learning goals. The importance of such
behaviours was echoed by many studies (Huh & Reigeluth, 2017; Kizilcec et al., 2017; Schunk,
1990). Huh and Reigeluth (2017) argued that goal setting is an iterative procedure in SRL that
was supported by overarching self-efficacy and self-motivation. Kizilec et al. (2017) highlighted
that goal seeing and strategic planning positively predict academic achievements. Further, the
results also revealed that with self-reflection, peer assessment rubrics and teachers’ feedback
provided, the students were willing to reflect on whether there were better ways to improve
their final projects, namely their English writing, storyboard and their own digital stories. It is
believed that self-reflection can be cultivated when reflection prompts and feedback from
teachers and peers are provided to engage students in self-reflection (Masui & De Corte, 2005;

6. Conclusion, Limitations and Future Studies

In the study, an innovative pedagogy, namely SRDST, was employed in primary school
students’ formal English learning. In order to understand its effectiveness, an instrument with
a factor analysis yielding three factors 1) forethought. 2) performance 3) reflection was
developed. The factors were similar to the ones identified in the literature (Zimmerman, 2002).
Based on the SRL framework for guiding teaching and learning, a curriculum with teaching
and learning resources were developed for implementing SRDST in English learning. The t-
test results suggested a significant improvement in the forethought phase where the students
believed they know where and when they can learn English most efficiently, they reckoned
that they can consult others when they are in need, and they possessed the autonomy to
achieve learning goals in their own ways. Besides, the students managed to monitor their own
progress as long as clear guidelines and guidance were given to them. Further, after they
finished the first draft of English writing and their first digital stories based on their own English
writing storyboards, they would reflect on whether there were better ways for refinement
according to the self-reflection rubrics as well as the feedback from teachers and peers. The
positive results revealed the feasibility of using SRDST to teach and learn English language.
Further, the SRDST has much potential in innovating teaching and learning. For example, this
pedagogy can be used in other disciplines, e.g. science, history, for steering the development
of curriculum resources. Besides, this learning strategy can be taught to young learners for
cultivating good learning habits, developing digital competencies, and preparing them for
becoming lifelong learners in a digital world.

Admittedly, this research may have some limitations due to the limited sample size. First,
the experiment time only lasted for 3 weeks and covered only one English unit due to the
school’s class planning and limited curriculum space. A longer duration with more suitable
topics to be covered would be ideal for further exploration and later a broader implementation.
Second, the current study was conducted in one primary school and the sample size only met
the minimum requirement of delivering an exploratory factor analysis (Bryant & Yarnold, 1995).
To better understand the relationship between students’ gender, interests and motivations, a
large sample size with participants from different schools was desired to examine the factor structure and access the effectiveness of this pedagogy and learning strategy. Third, the teachers' perspectives on the SRDST pedagogy were not clear. Further studies need to be done by organizing in-depth teacher and student interviews, class observation and content analysis on students' works.

References


Constructing a Natural Conversation Learning Activity to Improve Students’ English-Speaking Skills

Hong-Min TU\textsuperscript{a}, Chang-Yen LIAO\textsuperscript{b}, Te-Yang CHOU\textsuperscript{c}

\textsuperscript{a} Graduate Institute of Network Learning Technology, National Central University, Taiwan
\textsuperscript{b} Department of Hakka Language and Social Sciences, National Central University, Taiwan
\textsuperscript{c} Graduate Institute of Network Learning Technology, National Central University, Taiwan

*benjamintu1514@gmail.com

Abstract: As a medium of communication, language encapsulates four core competencies: listening, speaking, reading, and writing. Notably, speaking, which facilitates direct communication, is critical. Yet, Taiwan’s English pedagogy heavily emphasizes reading, inadvertently fostering a deficit in speaking skills among students. Addressing this, the current study proposes an English conversational practice system that adheres to language acquisition tenets, integrating 10-minute mini-lessons for English listening and speaking exercises. In these activities, students utilize a natural approach, engaging with the system-provided English conversational audio to refine their listening and speaking skills. The system, furnished with a recording function, empowers students to engage in self-paced English listening and speaking practice during their free time. The process of "listening to system audio, followed by speaking practice (recording), and re-recording" offers students corrective feedback from their recordings, enhancing their pronunciation, grammar, and fluency - vital elements of conversational competency. Additionally, the system incorporates a voice-based chatbot, facilitating situational speaking exercises based on the practice topics. The overarching goal of this study is to instill in students the skill to transfer their learning to diverse conversational contexts they might encounter, fostering habitual speaking to mitigate the disparity in the development of the four English language skills.

Keywords: conversation learning, mini-lessons, language acquisition, voice recording, natural approach

1. Introduction

The English education system in Taiwan has been overly focused on the "reading" skill (understanding English article content and grammar structures), which has resulted in Taiwanese students performing significantly better in "reading" skills than in "speaking" and "listening" skills (National Development Council of Taiwan, 2021). Our language skills are acquired through the acquisition, while "acquisition" is a part of the learning process, referring to learning non-deliberately or unconsciously (Mani, 2016). Children acquire their native language skills through constant exposure to conversations in daily life (Clark, 2018). Similarly, students’ speaking skills develop through listening to a large amount of external input in daily life and imitating what they hear and understand.

Based on the above, our aim is to develop an English conversation practice system for students that provides a conducive language learning environment and designs 10-minute mini-lessons. This will allow them to acquire English naturally through the imitation of the process of native language learning (Krashen & Terrell, 1983). This system enables students to engage in "self-paced English listening and speaking practice" or "situational conversation practice with a voice-based chatbot", using tablets or mobile devices during their free time.

Afterwards, students can listen to different conversation topics and practice listening and speaking independently through the "listening to system audio, followed by speaking practice (recording), and re-recording" mode. By utilizing recordings, students can improve
their pronunciation, grammar, fluency, intonation, and more, thereby increasing their confidence and courage to use English in spoken interactions (YanJu, Mei & Mohamed, 2017). Voice-based chatbots enable students to practice spoken language, meeting their self-paced needs (Chen, Vicki Widarso & Sutrisno, 2020), and are not limited by time or location (Han, 2020). Additionally, these chatbots provide students with an authentic language environment, contributing to enhancing their speaking skills (Huang, Hew & Fryer, 2022), ultimately forming the habit of speaking English.

2. System Introduction

In order to improve the current English education system in Taiwan, which finds it difficult to cultivate students’ speaking skills in English class, the design philosophy of this study system is based on the "Natural Approach" (Krashen & Terrell, 1983) and "Interest-Driven Creator Theory" (Chan et al., 2018), using the habit loop's “cuing environment” and "routine activities". This allows students to immerse themselves in a language acquisition environment constructed by the English conversation practice system, and by being exposed to significant amounts of English language input, cultivates students' speaking skills and develops the habit of speaking English.

The research team has compiled 36 English conversation learning topics, including socializing at school, family interaction, and time concepts. They have collected related English conversation content for each topic, which has been categorized into five levels of difficulty. So far, the system has collected approximately 2,000 English conversations.

In the mini-lessons, teachers can select an English conversation topic that matches their students’ English level and learning progress from the English conversation topic page. The blackboard page displays the English text and Chinese text (including Mandarin phonetic symbols) of the conversation content, and also allows for the playback of the English conversation audio. This enables students to practice speaking along with the English conversation content they hear.

This system enables students to engage in "self-paced English listening and speaking practice" or "situational conversation practice with a voice-based chatbot", using tablets or mobile devices during their free time. The voice-based chatbot will employ OpenAI as the conversation partner for students to practice speaking with. However, as OpenAI is based on text-based artificial intelligence, Microsoft Azure Speech Services will be utilized as the communication bridge between the learners and OpenAI. The process of situational conversation practice will begin with OpenAI presenting a topic related to the theme, and then the text will be converted to speech via Azure Speech Services and played back. Students will listen to the content and provide oral responses, which they will record. The recorded speech will be converted back to text using Azure Speech Services and sent to OpenAI. Through this process, students can apply the conversations they have practiced to real-life scenarios, thereby enhancing their speaking skills.

3. Preliminary Evaluation

3.1 Research participants

The research participants for this study were 55 students from a certain experimental primary school in Taoyuan city, Taiwan, ranging from grades three to six. Of the participants, 24 were boys and 30 were girls. Other participants included researchers and the English teachers from each class.

3.2 Preliminary Results and Conclusions

The objective of this study is to examine the impact of English listening and speaking exercises, implemented through mini-lessons and self-paced activities, on the oral proficiency
and habits of elementary students in grades three to six. Observations within the classroom setting have demonstrated that, with teacher guidance, students actively participate in English listening and speaking practice activities. Interestingly, after engaging in these mini-lessons, students show a growing willingness to communicate in English with their educators and classmates during the sessions.

Initial findings of this study suggest that English speaking and listening practice, delivered through mini-lessons and self-paced activities, positively impacts students' inclination and habits to communicate in English. As the research is set to continue over the full semester, we will maintain our classroom observations and accumulate data on students' midterm and final English grades. At the end of semester, we intend to carry out an English-speaking competency assessment, utilizing evaluative criteria from a preceding study on English speaking performance (West Hartford Public Schools, 2012). This assessment will encompass six dimensions: "accuracy," "comprehensibility and pronunciation," "fluency," "comprehension," "content," and "maturity of the language."

Acknowledgements

This study was funded by the National Science and Technology Council of Taiwan (112-2410-H-008-020-MY3) and by Research Center for Science and Technology for Learning, National Central University, Taiwan.

References

West Hartford Public Schools. (2012). *World Language Department’s middle school speaking rubric*.
Analysis of topic sentences of classification paragraphs and development of a diagnostic function

Kandai ISHIKAWAb*, Hidenobu KUNICHIKAb
\textsuperscript{a}Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan
\textsuperscript{b}Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan
*k.ishikawa@minnie.ai.kyutech.ac.jp

Abstract: In a paragraph, the topic sentence is an important component that clearly expresses the author’s argument and indicates the direction of the entire paragraph. It is difficult for beginners to learn English to write appropriate topic sentences. In order to solve the problem, we have focused on classification paragraphs and developed a diagnostic function of the topic sentences. As the result of an evaluation of the function, we have found that the function correctly diagnosed 70.6\% of appropriate topic sentences taken from text books for paragraph writing.

Keywords: Support in English Language Learning, Paragraph Writing, Topic Sentence, Classification Paragraph

1. Introduction

Paragraph writing is one of writing techniques for expressing opinions and arguments. In a paragraph, the topic sentence is the most important component that clearly expresses the author’s argument and indicates the direction of the entire paragraph. Moreover, a well-written topic sentence determines whether a paragraph can be written well. There are some conditions for writing topic sentences, and they differ according to types of paragraphs. In order to write paragraphs appropriately, it is necessary to understand such requirements, but beginners of English learning often lack sufficient knowledge. Therefore, we focus on supporting the writing of topic sentences.

There are some studies on assessing writing; e.g., (Su, Lin & Lai 2023). Although such systems will give learners feedback that is somewhat vague and abstract, they will not be able to generate concrete feedback according to types of paragraphs. In order to solve the problem, we are aiming to implement a diagnostic function of topic sentences as one of the functions of an English language paragraph writing support system. In this study, we focus on classification paragraphs, and have implemented a prototype of a diagnostic function of topic sentences (Ishikawa & Kunichika, 2023). This paper describes development and an evaluation of the diagnostic function.

2. Topic Sentences of Classification Paragraphs

Classification paragraphs classify things, that correspond to the topic of the paragraph, into several groups according to their characteristics or features. A topic sentence basically consists of two parts: a topic that indicates what the entire paragraph is about, and an assertion that expresses the author's argument or opinion on the topic. In this study, we have investigated the conditions of assertion in topic sentences of classified paragraphs by referring to textbooks on paragraph writing, and classified the conditions into three types. Also, we have categorized the forms of topic sentences into two types. The
three types of conditions differ in their applicability depending on the form of a topic sentence. The forms of topic sentences and the conditions of assertions are as follows.

(Form 1) The topic is appeared as the subject of a topic sentence.
The subject of a topic sentence corresponds to the topic, and is followed by a verb used for classification; e.g., "Popular music can be divided into three types."

(Condition 1) The verb of a topic sentence needs to mean classification.
This condition is applied to only Form 1. The topic sentence of a classification paragraph should indicate that the topic is classified into several groups. So, the verb of a topic sentence needs to mean classification. There are some frequent verbs such as "be divided into" and "fall into" (Kitao & Kitao, 1998).

(Condition 2) A criterion for classification needs to be provided.
The details of each category are presented in supporting sentences of a paragraph. Therefore, the number of classified categories needs to be indicated in a topic sentence based on the criterion or the number of categories. e.g., "divided into three types", and the other is to list the name of each category, e.g., "classified as A and B."

(Condition 3) The number of classified categories need to be indicated.
The number of classified categories needs to be indicated in a topic sentence (Donald et al.,1998). There are two ways to express the number of items: one is to indicate both the criteria for classification and the number of categories, e.g., "divided into three types", and the other is to list the name of each category, e.g., "classified as A and B."

3. Development of a Diagnostic Function of Topic Sentences

Our diagnostic function analyzes a topic sentence by referring to the result of word dependency analysis performed by CoreNLP (Stanford, 2022). Figure 1 shows the result of word dependency analysis of "Popular music can be divided into three types." Then, the function diagnoses if a topic sentence satisfies all the conditions. The diagnosis of topic sentences consists of two steps: identifying the form of a topic sentence and judging the conditions of the assertion of the topic sentence. First, for identifying the form of a topic sentence, the function confirms whether or not the main verb, which is shown as a root in the output from CoreNLP, is a be-verb and whether or not the subject is an expletive. Next, the function judges each condition of an assertion. For the judgment, the function extracts important words from the result of word dependency analysis performed by CoreNLP. Important words correspond to the main verb, a criterion for classification, and the number of classified categories. If the topic sentence is "Form 1", the function confirms whether or not the verb means classification by referring to a list of frequent verbs (e.g., Howard & Tracz, 1976; Donald et al.,1998) for condition 1. For condition 2, the function confirms whether or not the criterion for classification needs to be provided.
not a word corresponding to the object of the topic sentence can be used for a criterion of classification by referring to a list of frequent words which express criteria (e.g., Howard & Tracz, 1976; Donald et al., 1998). For condition 3, the function confirms whether or not there are numeral modifiers for the words which express criteria or whether or not the names of categories are listed. Finally, the function outputs the result of the diagnosis.

When an input sentence satisfies all the conditions, the function shows the format of the topic sentence, the verb which means classification, the criteria for the classification, and the number of classified categories. Figure 2 shows an example of the result of an acceptable sentence. If an input sentence does not satisfy the conditions, the function shows the conditions that are not satisfied.

<table>
<thead>
<tr>
<th>INPUT</th>
<th>Popular music can be divided into three types.</th>
</tr>
</thead>
</table>
| DIAGNOSTIC RESULT | *The topic sentence entered is appropriate  
Topic Sentence Form: Topic as the subject  
Verbs that imply classification : divide  
Criteria for classification : type  
Number of classification : three |

Figure 2. An Example of the Result of an Acceptable Sentence (Ishikawa & Kunichika, 2023)

4. An Evaluation of the Diagnostic Function of Topic Sentences

As an evaluation, we prepared 17 topic sentences of classified paragraphs taken from eight textbooks (e.g., Howard & Tracz, 1976; Donald et al., 1998; Oshima & Hogue, 2017) on paragraph writing, and used the sentences as input for the diagnostic function. As the result, the system correctly diagnosed 12 (70.6%) of the 17 sentences as appropriate, while the other 5 sentences were judged as unacceptable sentences. We have found two patterns of sentences that were misdiagnosed: one is omitting the criteria for classification and the other is listing categories in clauses or sentences connected by a conjunction.

5. Conclusion

In this study, we have developed a diagnostic function of the topic sentences for classification paragraphs. Also, this paper has presented an evaluation of the function. As the result of the evaluation, we have found that the function correctly diagnosed 70.6% of appropriate topic sentences taken from textbooks for paragraph writing. In the future, we plan to evaluate the function by using sentences including inappropriate topic sentences.

References

Construction of an English Grammar Quiz Recommendation System Using Explanation by a Knowledge Map

Kensuke TAKII*, Naomichi TANIMURAa, Brendan FLANAGANb & Hiroaki OGATAc

aGraduate School of Informatics, Kyoto University, Japan
bCenter for Innovative Research and Education in Data Science, Kyoto University, Japan
cAcademic Center for Computing and Media Studies, Kyoto University, Japan

* kensuke.takii96@gmail.com

Abstract: Many systems to assist in learning English grammar have been developed in the field of learning material recommendation systems (LMRSs). Compared with those based on knowledge models, recommendations based on data tend to cause the cold-start problem, and it is said that explainable LMRSs may be able to enhance learners’ motivation. In our research, we propose an explainable English grammar quiz recommendation system using a knowledge map to support students’ learning of English grammar with trust in the system and motivation. The learning effect of the explanation of the system was evaluated in an experiment, in which 349 high school students in Japan participated. This experiment showed that there was little learning effect of the explanation, but the system reliability and motivation improved by the explanation. The limitation and our future work regarding the validity and learning effects of the system are also indicated.

Keywords: English grammar learning, e-learning recommender, knowledge map, learning path, explainable recommendation

1. Introduction

There has been increasing need for learning material recommendation systems (LMRSs), because of the difficulty for learners to select appropriate materials due to the increasing options (Liu, 2019). Although existing LMRSs can be classified into systems based on educational data and those based on domain knowledge models, the former is more likely to cause the cold-start problem for new users than the latter (Sun et al., 2019). Additionally, recent studies have developed explainable LMRSs such as Takami et al. (2022), which can explain the rationale of the recommendations. Since feedback from systems improves learners’ motivation and achievement (Duffy and Azevedo, 2015), explainable LMRSs based on knowledge models may improve the learning effect and motivation in learning.

In the field of LMRSs, many systems to assist in learning English grammar have been developed (Fang et al., 2018). English is an important second language in many regions of the world, and the grammatical structures present in various languages pose challenges for reading comprehension in foreign languages (Alderson et al., 1984). According to previous research, teaching grammar items in a specific order is necessary to gain maximum educational effect (Izumi and Isahara, 2004). Hence, the order in which grammar items should be learned is vital in learning English as a second language (ESL).

In our research, we propose an explainable English grammar quiz recommendation system using a graph structure called a knowledge map. We developed an English grammar knowledge map as a directed graph structure that can represent English grammar items, their relations, and the order in which they should be learned. Besides, the recommendation using the knowledge map can explain its rationale by tracking learners’ learning logs and recommended quizzes. In addition, we implemented the system into an existing web-based learning system and conducted an experiment targeting Japanese high school students who
are studying ESL to verify the effect of the explainable recommendation in learning English grammar. Our research focuses on the explanation of the rationale of the recommendations and the following research question is set: Does the explanation of the rationale of the recommendation for learning English grammar have a learning promotion effect?

2. Proposed Approach

2.1 English Grammar Knowledge Map

We developed an English grammar knowledge map, which represents English grammar items, the relations between them, and the learning order as a graph structure. We referred to CEFR-J Grammar Profile (Ishii and Tono, 2018), as a list of English grammar items. It includes 501 grammar items, and each item is classified into the CEFR (Council of Europe, 2011) level. Each item in the profile has a regular expression, which is used to judge if the item is used in a certain sentence. For example, the sentence “I am a student.” includes the grammar item “Personal pronoun nominative (I)+be: I am”, and the system can automatically determine that the sentence includes the grammar item by using the regular expression.

The knowledge map was generated based on the method by Flanagan et al. (2019): The nodes representing grammar items in CEFR-J Grammar Profile are connected by directed weighted edges, which represent the similarity of the regular expressions of the items and their proper learning order. The order was based on “English Grammar/Syntax Vintage” (Shinoda and Yoneyama, 2010), a workbook used for ESL learning in Japanese high schools. The knowledge map is structured as a directed maximum spanning tree by pruning the edges.

2.2 Explainable Quiz Recommendation with Knowledge Map

To recommend quizzes, our system first judges if each grammar item is learned or misunderstood by learners. The system creates two grammar sets using learners’ learning logs: The first is Incorrect Grammar Set (IGS), including grammar items in quizzes that the user has answered incorrectly, and the second is Unanswered Grammar Set (UGS), including grammar items in quizzes that the user has not yet answered. Then, the system recommends quizzes based on the following three conditions regarding the number of grammar items included in the quizzes. The higher these numbers, the more highly the quizzes are recommended; however, these 3 numbers are used in order of priority from 1 to 3. Besides, if there exists a directed edge from A to B in the knowledge map, we call A “a prerequisite” of B.

1. The number of grammar items that are prerequisites for other grammar items in IGS, and that are also included in IGS
2. The number of grammar items that are prerequisites for other grammar items in IGS, and that are included in UGS
3. The number of grammar items that are not prerequisites for other grammar items in IGS, but which themselves would be included in IGS

In generating an explanation in the proposed system, when the system recommends grammar quizzes, the grammar items are distinguished by the following criteria: whether the learner has incorrectly answered quizzes including these items before, or the learner has not answered these quizzes yet. Furthermore, the system notifies the learner whether a certain grammar item is a prerequisite to be known for another item that the learner has misunderstood, i.e., the learner has incorrectly answered quizzes including the grammar item.

2.3 User Interface of the Platform

As an interface for answering quizzes, we adopted BookRoll (Flanagan and Ogata, 2018), a multifunctional e-book reader system, and registered English grammar quizzes in a workbook being used in Japan for learning English. BookRoll has many functions to realize e-learning environments; one of which is a quiz function, with which learners can answer quizzes
registered to materials. We made e-books for answering the grammar quizzes, by editing the contents of the workbook so that each page has one quiz. Users can solve the quizzes on BookRoll and register whether their answer for each quiz was correct or incorrect.

The user interface of our recommendation system provides learners with information on (1) quizzes recommended by the system, (2) links to open the quizzes on BookRoll, and (3) explanations of why the recommendation was made for the learner. The interface displays 5 recommended quizzes, each of which has at most 3 types of explanations for why the quiz was recommended. These explanations clarify what kind of grammar items the user can learn with the recommended quizzes.

3. Experiment

3.1 Method Overview

We conducted an online A/B test, targeting high school students in Japan who are learning ESL by the curriculum in Japan. The target students were divided into an experimental group and a control group. Students in the experimental group used a recommendation system that displayed the explanation of the recommendation rationale, and students in the control group used a system that did not. Furthermore, we set the following 4 perspectives from which the effect of the explanations according to the RQ shown in the Introduction section:

- The number of recommendation usage: How highly did the students engage in learning?
- Scores of the pre- and post-test: Change of the learners’ performance
- Answers for the poll: Willingness to learn with the system and trust in it

3.2 Settings

The experiment targeted 349 first graders in 9 classes at high school (about 40 students per class, the students who did not agree to participate were excluded). The participants were divided into the experimental group (4 classes, 152 students) and the control group (5 classes, 197 students), both of which used the recommendation system while learning for 14 days. This experiment was conducted during a winter vacation, and the students were encouraged in doing the learning activity with the system as a review, after learning English grammar with a paper workbook. This online learning activity was done with BookRoll, and 219 grammar quizzes were registered to it. Table 1 shows an example of grammar quizzes.

Table 1. Examples of the quizzes registered to BookRoll

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>We are (   ) (   ) have you in our team.</td>
<td>lucky to</td>
</tr>
<tr>
<td>(Write an appropriate word to match the Japanese sentence.)</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Experimental Procedure

Step 1. The English teacher explains how to use the system to the participants.
Step 2. Before starting the learning activity, the participants take an English grammar pre-test.
Step 3. The participants solve some quizzes with BookRoll.
Step 4. The participants work on the quizzes using the recommendation and BookRoll.
Step 5. After ending the learning activity, the participants take an English grammar post-test.
Step 6. The participants answer the poll.

The contents of the pre and post tests were 5 grammar quizzes, which were randomly selected from the quizzes registered to BookRoll. The contents of both the quizzes were same. The poll was intended to an evaluation of the motivating effects of the system from the
viewpoints of trust and motivation in learning. This poll consists of 11 questions with a 5-Likert scale and 2 descriptive questions. For measuring the trust and motivation in using the system, we adopted Hoffman et al. (2018)’s “Trust Scale Recommended for XAI” and Keller (1987)’s ARCS model. The perspective of the questions is shown in the Table 3 in Result section.

4. Result

4.1 Usage of the System

Table 2. Numbers of students who used the system

<table>
<thead>
<tr>
<th>Participant</th>
<th>Recommendation</th>
<th>BookRoll</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>With explanation</td>
<td>152 (16 (10.5%))</td>
<td>10 (6.6%)</td>
<td>26 (17.1%)</td>
</tr>
<tr>
<td>Without explanation</td>
<td>197 (18 (9.1%))</td>
<td>10 (5.1%)</td>
<td>28 (14.2%)</td>
</tr>
<tr>
<td>Total</td>
<td>349 (34 (9.7%))</td>
<td>20 (5.7%)</td>
<td>54 (15.5%)</td>
</tr>
</tbody>
</table>

Table 2 shows the numbers of students who used both the recommendation system and BookRoll, who accessed the recommendation interface, and who used BookRoll only. This table indicates that about 10% of the participants used the recommendation and that about 6% of them used BookRoll only, without using the recommendation.

4.2 Result of the Poll

Table 3. Result of Mann-Whitney’s U test on the questions in a 5-Likert scale of the poll

<table>
<thead>
<tr>
<th>Question</th>
<th>Perspective</th>
<th>Group</th>
<th># of Users</th>
<th>Mean</th>
<th>Std.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Trust</td>
<td>With explanation</td>
<td>14</td>
<td>3.29</td>
<td>0.61</td>
<td>P=0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>3.06</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>Trust</td>
<td>With explanation</td>
<td>14</td>
<td>3.64</td>
<td>1.01</td>
<td>P=0.58</td>
</tr>
<tr>
<td>(predictability)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>3.06</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>Trust</td>
<td>With explanation</td>
<td>14</td>
<td>2.93</td>
<td>1.00</td>
<td>P=0.69</td>
</tr>
<tr>
<td>(reliability)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.78</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>Trust</td>
<td>With explanation</td>
<td>14</td>
<td>2.86</td>
<td>1.17</td>
<td>P=0.64</td>
</tr>
<tr>
<td>(efficiency)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.61</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>Trust</td>
<td>With explanation</td>
<td>14</td>
<td>2.57</td>
<td>1.16</td>
<td>P=0.98</td>
</tr>
<tr>
<td>(endorsement)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.56</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>Motivation</td>
<td>With explanation</td>
<td>14</td>
<td>2.57</td>
<td>0.94</td>
<td>P=0.79</td>
</tr>
<tr>
<td>(alert)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.50</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>Motivation</td>
<td>With explanation</td>
<td>14</td>
<td>3.14</td>
<td>1.10</td>
<td>P=0.56</td>
</tr>
<tr>
<td>(relation)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.94</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>Motivation</td>
<td>With explanation</td>
<td>14</td>
<td>4.07</td>
<td>0.73</td>
<td>P=0.61</td>
</tr>
<tr>
<td>(confidence)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>4.00</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Q9</td>
<td>Motivation</td>
<td>With explanation</td>
<td>14</td>
<td>2.79</td>
<td>1.12</td>
<td>P=0.66</td>
</tr>
<tr>
<td>(satisfaction)</td>
<td></td>
<td>Without explanation</td>
<td>18</td>
<td>2.61</td>
<td>0.98</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 summarizes the result of the poll. It shows that from both the perspectives of trust and motivation, the impression of the students in the experimental group was more favorable than that of the students in the control group. In terms of Trust, there are differences of more than 0.2 points in the mean scores in the items of trust and efficiency. In terms of Motivations, there are differences of more than 0.2 points in the items of relation. In the free-text items, the answers regarding the efficiency in review and finding weak points were the most common response which was favorable for the recommendation system. On the other hand, as the response that was not favorable, we found the answers that doubted the correctness of the diagnosis by the system and found selecting the quizzes troublesome. In addition, some students found the system interface inefficient.
4.3 Result of the Pre- and Post-test

Figure 1 shows the distribution of the results of the pre- and post-test and the change in the scores. The scores of the post-test were higher than those of the pre-test, but this difference was not statistically significant. Comparing the experimental and control groups, the scores of the pre- and post-test and the change in the scores did not indicate significant differences.

5. Discussion

5.1 Learning Promotion Effect by the Explainable Recommendation

In the post-poll, there were relatively large differences on the items that asked about trust, predictability, reliability, and efficiency in the Trust items. This implies that the explanation of the recommendation rationale improved the reliability of the system as perceived by the users. On the other hand, the item that asked about endorsement showed less difference. Considering that many answers for the free-text questions referred to the difficulty in using the system, the use of the system itself was a burden in learning. Also, the reliability of the explanation would be less influential on the system acceptance than the system’s user-friendliness. In the question about Motivation, there were relatively large differences on the items about relation and satisfaction, which implies that the display of explanation of the recommendation rationale made users feel the relationship between their proficiency and learning contents proposed by the system. Furthermore, in the free-text questions, some users in the control group doubted the reliability of the diagnosis by the system, and such answers could not be found in the experimental group. This implies that it is difficult for users to judge the correctness of the recommendation without an explanation of the rationale.

The result of the grammar tests shows no significant differences between the two groups, and the post-test scores were higher than the pre-test scores but not significantly different from them. This implies that the student did not use the system often enough for the difference in the condition to influence the scores. Also, the number of quizzes in pre and post tests may be so small that no significant differences in the grammar tests were found.

5.2 Limitation

Our system was used infrequently regardless of conditions. We argue that the main reasons were lack of usability, and the burden in reading the explanation. When the students solve the quizzes with BookRoll and the proposed system, they must switch their screens many times. Second, if reading the explanation requires long time, it may be the users’ burden to propose the explanation. Next, our system does not consider the level of importance of each grammar item in learning. Each quiz includes several grammar items, some of which are necessary to solve it, but others are not. Without considering the importance of each grammar item, the system might recommend quizzes that include items not important for solving them.
6. Conclusion and Future Work

Our experimental result did not show a learning effect from the explainable recommendation but implied that the system reliability and motivation were improved. Therefore, the answer to our research question was that the explainable recommendation improved the users’ learning in the aspects of (1) reliability and motivation and (2) the frequency of the system use.

Our future work will focus on the validity and learning effects of the system. First, we should verify the accuracy of the knowledge map and the learning effects by considering learning paths. Second, the sufficient result has not been gained about whether the explanation of the recommendation rationale has a learning effect. The burden of using the system and understanding the explanation should also be investigated.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, and NEDO JPNP20006.

References


Educational System of English Tense for Japanese Learners by Forming Temporal Constraints on Tense

Haruto NAGATAa*, Tatsuhiro KONISHIb, Makoto KONDOb,c

*Graduate School of Integrated Science and Technology, Shizuoka University, Japan
b,cFaculty of Informatics, Shizuoka University, Japan

nagata.haruto.18@shizuoka.ac.jp

Abstract: In this study, we developed an English-tense learning and education system that provides training in constructing tense constraints from given sentences. Temporal constraints refer to the condition that the correct choice should be met in English tense-choice questions. In this study, we organized the patterns of temporal constraints and guided learners to assemble them based on these patterns. We also introduced a visual representation to enhance the intuitive understanding of temporal constraints. We conducted an experimental evaluation of the proposed system to investigate its learning effectiveness. Additionally, in the experimental evaluation, we developed a system using graphic representation and a system without graphic representation and investigated the differences between them.

Keywords: Learning Support System of English grammar, Tense in English, Temporal Constraints, Graphic Representation

1. Introduction

In English language learning among Japanese native speakers, errors in verb tense largely depend on knowledge outside the text, including real world knowledge. As a result, correcting verb tense and aspect errors has been suggested to be difficult (Lee, 2008).

In this study, we focused on verb tense and aspect in the context of English tense multiple-choice exercises in which learners are required to select the appropriate verb form from a set of choices to fill in a blank. The process of solving English tense problems can be considered as (1) constructing constraints about the word that should fill in the blank in the problem sentence and (2) determining whether each choice satisfies all the constraints. We call these constraints “temporal constraints.” In English tense problems, temporal constraints are frequently encountered and are important among those constraints. Several errors of English are believed to occur in the (1). Therefore, we developed a learning and educational support system that trains learners to construct temporal constraints from the information contained in problem statements. To assist in the construction of tense constraints, we break down temporal constraints into patterns. By conducting the learning process of constructing temporal constraints based on each pattern, we aim to improve the accuracy of answering the questions and enhance the ability to explain the cognitive reasoning underlying the answers.

As a learning method for constructing temporal constraints, we envision an interactive process between learners and an educational support system that guides them in following (1) and (2). In Step (1), the system guides learners to construct temporal constraints from the problem text based on the temporal constraint patterns. In Step (2), the system allows learners to apply constraints to each choice and determine whether the constraints are satisfied. The constraints can be expressed in text format. However, we are afraid that some types of temporal constraints in text are difficult to intuitively apply to the choices in the Step (2). Therefore, we introduce a graphical-representation method to express these constraints.

The reasons for introducing the use of graphical representation are as follows.
Problems with English tenses require an understanding of the temporal relationships between events in English sentences. In general, time is recognized as a one-dimensional space from the past to future. Thus, by representing the timeline as a one-dimensional space and displaying a diagram that positions events on the timeline, the temporal relationship between the events can be clarified.

In this study, to verify the effectiveness of this graphical representation, we constructed and conducted an experimental evaluation of both a system with and without graphic representations. The following results were obtained for both systems. Before and after receiving support from the systems, a significant improvement was observed in both the accuracy of problem solving and the quality of explanations for the answers. By using systems to practice exercises repeatedly, learners can internalize the methods for constructing temporal constraints without relying on the systems.

2. Related works

Many educational systems of English grammar for Japanese learners have been developed (Le,2011)(Ozaki,2004)(Ozaki,2007). However, these systems don’t aim to teach usage of verbs. The misuse of verbs such as tense and aspect is a typical error category for nonnative speakers of English. Correcting errors related to tense and aspect, which are directly related to the English verb tense, is considered difficult. This is because tenses and aspects heavily depend not only on information within the target sentence, such as adverbs and coordinating conjunctions, but also on the global context, such as knowledge of the real world outside the target sentence (Lee,2008). In the KJ Corpus, which annotates grammatical errors and performs syntactic analysis, article errors are the most frequent, followed by errors in singular/plural nouns, prepositions, and verb tenses.

Various methods have been proposed to support learning and education through the automatic detection and correction of errors(Page,2003)(Burstein,2003)(Liao,2016). However, there have been few developments in the creation of systems that specifically focus on thought processes involved in solving problems. As an educational system that focus on thought process, the system establishes thought process for arithmetic problems by having students select and arrange sentences from a provided collection of short sentences and pose questions in an appropriate order(Hirashima,2007). A system also focuses on developing thought process of programming using code puzzles(Ito,2020). Code puzzles are having students form programs by arranging statements that are ordered randomly. These systems are designed to have students form something in order to focus on thought process in solving problems. But they aren’t systems for English tense. In terms of learning of English tense, what should be formed by students? We think forming temporal constraints is effective for students to learn thought process of English tense problems. So we developed educational systems that teach students how to form the temporal constraints required for correct answer choices in English tense choice questions.

As a method for teaching students how to form temporal constraints, we focused on using graphical representations. Some studies use them to assist Japanese in understanding English tenses(Tanaka, 2018)(Miura, 2010). However, these graphical representations are not intended for application in learning English tenses; therefore, we thought an extension of the graphical method is necessary. In this study, we developed graphical representations suitable for learning and educational support, as described in Section 3.2. We developed learning and educational support systems to help learners judge whether each choice is valid by comparing the temporal constraints and events represented by each choice graphically.

3. Basic Ideas

3.1 Temporal constraints

We analyzed the relationship between the validity of choices and temporal constraints for 24 English tense choice questions from standard university entrance exam questions in Japan, including past exam questions from the National Center for University Entrance Examinations.
(2000-2018) and commercial study materials in Japan (Nakao, 2016). Consequently, we found that temporal constraints can be classified into the following three types:

1. Constraint by specific words or phrases: “already”, “by + time phrase”, and other specific adverbial or prepositional phrases determine the tense and aspect of the verb they modify.
2. Constraint by tense agreement: When the main clause is in the past tense, the subordinate clause also becomes past tense or past perfect tense.
3. Constraint based on the relationship between events: Constraints based on the context of the problem and common sense of the real world determine tense and aspects of verbs.

The systems that we developed allow learners to form these constraints in their sentences. For a system that uses graphical representations, the generated constraints can be incorporated into graphical representations.

3.2 Extending Graphical Representation

Based on the above problem analysis, we organized the necessary components for expressing the temporal constraints in a diagram as follows:

i) Event node: The element representing an event is expressed as a node, and the aspect of the event is expressed by the node type. These are shown in Figure 1.

ii) Timeline: We express the timeline using a horizontal arrow.

iii) Time of utterance: This is the time point when the problem statement was uttered.

In the timeline, the right side represents the future and the left side represents the past with respect to the point when the problem statement is uttered. For each event in the problem statement, we set the following components: i) ii) iii) For example, the graphical representation of the sentence “By the time the 2002 World Cup was held, soccer had already become a leading sport in Japan.” is shown in Figure 2.

3.3 Learning process in the system

The learning process in the system using graphical representations is shown in Table 1. On the other hand, system without using graphical representations, formation of graphical representations of temporal constraints and problem statement containing the choice of words and comparison of them are omitted, as described in Section 3.3.1.

<table>
<thead>
<tr>
<th>Step</th>
<th>Learning process in the system using graphical representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>For each of English question, the learner selects the validity of the answer with “O” for valid and “X” for invalid. If learners are unsure at the current stage, they can select “?”.</td>
</tr>
<tr>
<td>2</td>
<td>After considering the temporal constraints of the problem in a sentence using a fill-in-the-blank template, we form them as graphical representations.</td>
</tr>
<tr>
<td>3</td>
<td>Represent the problem statement containing a choice of words using graphical representations.</td>
</tr>
<tr>
<td>4</td>
<td>Compare the graphical representation of the temporal constraints created in Step 2 with the graphical representation of the answer choices created in Step 3 and determine whether the answer choice satisfies the temporal constraints.</td>
</tr>
<tr>
<td>5</td>
<td>The learner selects which part of the choice graphical representations they compared to graphical representations of temporal constraints. This helps eliminate random answers, estimates errors and reminds learners that they should focus their attention.</td>
</tr>
<tr>
<td>6</td>
<td>Step 3 to Step 5 are repeated for each choice.</td>
</tr>
<tr>
<td>7</td>
<td>After completing the comparison using graphical representations, the validity of each choice is selected as an answer by choosing “O” for valid or “X” for invalid.</td>
</tr>
</tbody>
</table>
4. Implementation

In this study, we constructed systems and implemented 10 problems that contained two types of constraints as a prototype: constraints by specific words or phrases and constraints by tense agreement. The constructed systems followed the learning flow described in Section 3.3. Systems are implemented using the C# language in Visual Studio 2022.

4.1 Forming and comparing graphical representations of temporal constraints with graphical representations of each choice

Temporal constraints were created as text by filling in the fill-in-the-blank template, as shown in Figure 3. Constraint types are displayed on the menu as guidelines to make learners aware of the pattern of constraints. By following a certain abstract pattern, learners learn to construct temporal constraints. If the learner makes a mistake, the system provides advice tailored to the blank space to help the learner arrive at the correct answer.

4.2 Forming and comparing graphical representations

Learners visualize the problem statement, including temporal constraints and each choice. They form graphical representations by specifying the temporal constraints and aspects of the node, as indicated by the red boxes in Figures 4 and 5. By visually comparing them, learners can judge whether each choice satisfies temporal constraints.

5. Experimental Evaluation

5.1 Hypothesis of the experiment

In the experiment, we validated the following experimental hypotheses. **Hypothesis 1:** Prompting learners to consider temporal constraints improves the accuracy of
their responses to questions and increases the quality of their explanations for judging the validity of their choices.

Hypothesis 2: Repeated practice exercises will help establish a consistent approach to judge the validity of each choice.

5.2 Experimental procedure

Participants were divided into two groups: Group A used a system with graphical representations, and Group B used a system without graphical representations. Five Japanese university students were recruited for each group and were given 10 English tense choice questions. Experimental procedure was performed according to Section 3.3 above. From the data obtained Step 1 and Step 7 in Section 3.3, aggregate how the reasons for each choice of “ألعاب”, “×”, and “?” changed. Thus, it can be said that experimental Hypothesis 1 is supported if the accuracy rate improves and the justifications for the validity of each choice provided by the learners become clearer, or if there is a reduction in ambiguous reasoning. In addition, if the reasoning for each choice in the system before learning becomes clearer through the repeated practice of problem-solving, Hypothesis 2 is supported.

5.3 Results

5.3.1 The correct answer rate

The correct answer rates for the questions in Groups A and B are shown in Figures 6 and 7. Here, the correct answer to a question means that all choices for that question are correctly answered with “ألعاب” or “×”. In both groups, “the correct answer rate improved significantly before and after using systems” (result1).

![Figure6. Accuracy of the Problems (group A)](image)

![Figure7. Accuracy of the Problems (group B)](image)

5.3.2 Changes in reasoning before and after system learning

We scored the validity reasoning for each choice before and after learning with systems as follows: Items that provide clear reasoning (2 points). Items that describe the type of temporal constraints, but do not provide specific measures to address the issue (1 point). Items that are clearly ambiguous, such as “I thought it seemed like it.” (0 point).

The total number of choices for the 10 questions was 165. Assuming that all 165 justifications for the choices are clear (2 points), the percentage of clarity in justifications before and after system learning for each group is shown in Figures 8 and 9.

In the justifications after system learning, it is clear that both Group A and B have clearer justifications than before system learning for all 10 questions. Therefore, we find that “using the system improves quality of explanations for the reasoning of the validity of the choices” (result2). Thus, the Experimental Hypothesis 1 is supported by the result1 and the result2. Furthermore, the changes in learners’ understanding through repeated practice can be seen from the score changes in questions 1-10 before system learning in Figures 8 and 9. As a result, “gradual improvement can be observed as the number of questions increases” (result3). Although this experiment used only 10 questions, we can expect sufficient
improvement if we increase the number of questions. Therefore, we can say that the Experimental Hypothesis 2 is supported by the result 3.

6. Conclusion

In this study, we developed learning education support systems that teach learners to construct temporal constraints based on information on English tense-choice problems. The experimental results showed that we could get “the correct answer rate improved significantly before and after using systems”(result 1), “using the system improves quality of explanations for the reasoning of the validity of the choices”(result 2) and "gradual improvement can be observed as the number of questions increases”(result 3) in Section 5.3. However, based on their responses to the questionnaire, some students noted that the graphic representations were effective. Therefore, we plan to conduct experiments to assess the effectiveness of graphics tailored to the learners’ English proficiency levels.

References

**Exploring the use of chatbot to promote online EFL students’ behavioral, cognitive, and emotional engagements**

Xinyi LUO\(^a\), Weijiao HUANG\(^a\)*, Khe Foon HEW\(^a\), Chengyuan JIA\(^b\) & Xiangjie CAO\(^c\)

\(^a\) Faculty of Education, The University of Hong Kong, Hong Kong SAR, China  
\(^b\) College of Education, Zhejiang University, Hangzhou, China  
\(^c\) College of Foreign Languages, Northwest University of Political Science and Law, Xi’an, China  
*wjhuang1@connect.hku.hk

**Abstract:** The implementation of online learning is often beset with the challenge of student disengagement, especially in online language learning where active learner practice is expected. To address this problem, we explored the use of a chatbot to support students with their English as a Foreign Language (EFL) listening learning in an online undergraduate course. The Self-determination theory was adopted to design the chatbot. Students were invited to interact with the chatbot for ten days to practice their English listening skills. A mixed-methods design was used to measure students’ behavioral engagement, cognitive engagement, and emotional engagement. The results indicated students’ positive online learning experiences with the chatbot. We also conducted follow-up interviews with students to determine their perceptions of learning with the chatbot.

**Keywords:** Self-determination theory, chatbot, online learning, language learning

---

1. **Introduction**

Over the past few decades, online learning has been viewed as a weaker alternative, and often stigmatized as a lower-quality form of education, compared to traditional face-to-face instruction (Hodges et al., 2020). However, due to the disruption caused by Covid-19, many countries are forced to switch to online learning (Barron Rodríguez et al., 2020). Educators had to employ online learning approach during the pandemic to sustain school education. Today, despite the end of the pandemic, the increasing of online learning has exceeded our expectations and grown in importance. Nonetheless, the persistent issue of student disengagement in online learning should not be overlooked (Dumford & Miller, 2018). Student disengagement refers to students’ low participation or interest in the learning activities, which negatively impacts their online learning performance. This problem also exists in online language learning (Ji et al., 2022), where students need to master the language skills by actively practicing the specific skill. Previous studies have identified several challenges that affect students’ engagement in online language learning, such as the lack of autonomy (Jiang & Peng, 2023), insufficient interactions (Tran, 2018), and inadequate teacher instruction (Indartono, 2019). One way to address this issue is to provide students with chatbots in their online language learning (Hew et al., 2023).

Chatbots, dialogue systems that interact with users via text or voice, are used primarily for language learning in education, particularly English learning. Previous research has demonstrated the positive effects of chatbots as conversational partners for students' language learning, with experimental groups showing significant performance improvements (e.g., Jeon, 2021). While some studies (e.g., Hew et al., 2023) highlight the potential of chatbots to enhance student engagement, others (e.g., Liu et al., 2022) found no significant impact on student interest in English reading activities. Given these mixed results, further research is needed to determine the effectiveness of using chatbots in language learning.
Existing studies (e.g., Yang et al., 2022) have primarily focused on chatbot applications in speaking, reading, and writing, with limited empirical research on the use of chatbots to improve students' EFL listening comprehension.

In this study, we used a chatbot as an interlocutor for EFL listening learning in an online listening course. The present study mainly aimed to investigate the impact of the chatbot on student engagement. Student engagement has broadly been described as a multidimensional concept that encompasses three main domains: behavioral, cognitive, and emotional engagement (Cooper, 2014). Behavioral engagement refers to students' participation in learning and academic assignments (Fredricks et al., 2004). Cognitive engagement concerns the level of students' desire to go beyond the course requirements in the learning (Fredricks et al., 2004). Emotional engagement was described as students' affective responses, like interest, satisfaction, and sense of anxiety (Fredricks et al., 2004). Additionally, we explored suggestions from students to further optimize the effectiveness of the chatbot in online language learning. This study is guided by the following research question: **What is the effect of the chatbot on students' engagement (behavioral, cognitive, and emotional engagements) in online English listening learning?**

2. Design and development of the chatbot

In this study, we designed the chatbot based on self-determination theory (SDT) aiming to engage students in online English listening. Engagement is considered an observable outcome of motivation (Reeve, 2012). SDT is a widely used theory for motivational development and psychological needs to deepen understanding of how and why individual behavior occurs (Deci & Ryan, 1985). SDT assumes three vital and inherent human needs that drive action: Autonomy, Competence, and Relatedness (Ryan & Deci, 2000). The need for autonomy is concerned with the desire of people to be causal agents and to have volition in their behavior (Ryan & Deci, 2000). More specifically, students' demand for autonomy can be satisfied if they can freely choose to make an extra effort or not. Competence pertains to whether one's pursuits or learning can be effectively mastered (Helme & Clarke, 1998). In the educational setting, students feel more competent when they can address the challenges of assigned tasks. Relatedness can be defined as students' desire to feel associated with others and to own a sense of belonging (Ryan & Deci, 2000). In teaching and learning, students have an imperative desire to interact with their peers and instructors. Figure 1 shows the chatbot interface.

![Chatbot interface.](image)

The autonomy in students' learning was achieved by an optional learning journal activity in the chatbot system, where students could reflect on their daily learning experiences. The chatbot prompted students to express their learning difficulties or reflections, by saying "How's your learning today? Any difficulties?". We considered students' journal writing as the indicator of their autonomy because they controlled the learning behaviors (deciding whether or not to write and what materials to reflect upon).

Competence was attained through three approaches: instructor-designed explanation videos, targeted feedback, and task instruction. First, the chatbot would deliver an instructor-
designed video that included the key points of each listening practice, helping students to comprehend the listening materials. Second, the chatbot would provide targeted feedback in terms of students’ input. For example, if students misspelled a required word, the chatbot would respond with target feedback, such as “The spelling of this word is not complete.” Third, the chatbot could provide step-by-step instructions on each listening task during the interaction. This aimed to alleviate any confusion that students may encounter during the learning process and enhance their competence to master the listening skills. For example, the chatbot introduced the learning objectives and guided students to check learning content by saying “You may check the explanation video below.”

To enhance relatedness, the chatbot used communication indicators such as emojis and encouraging prompts for human-like interaction. Emojis, which are widely used in digital communication (Boutet et al., 2021), can help the chatbot to express emotional responses. For example, the use of the celebratory emoji in a message, “Well done 🙂”, could help the chatbot convey more friendly features to students. In addition, encouraging messages were used to foster a positive learning environment. When students made a mistake, the chatbot would respond with, “Please try again. I believe you can do it!”

3. Research Design

Our target participants were 41 year-two undergraduate students at a large public university in northwestern China. The present study was conducted in a compulsory English course. The whole intervention lasted two weeks, and students needed to interact with the chatbot in ten days to practice English listening. Ethical approval to carry out the study was obtained from the author’s university and consent was received from all participants. Table 1 summarizes the quantitative data collection methods used in this study. To further examine students’ perceptions of using the chatbot, individual semi-structured interviews were conducted. Twelve participants attended the interviews voluntarily. A sample interview question was “How do you think about your learning with the chatbot?”

Table 1. Data collection methods

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Quantitative data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral engagement</td>
<td>Students’ listening task completion rate, which refers to the number of students who complete the required task to the total number of students. Students learning records were collected.</td>
</tr>
<tr>
<td>Cognitive engagement</td>
<td>Students’ learning journal completion rate. This is an optional learning record.</td>
</tr>
<tr>
<td>Emotional engagement</td>
<td>A 5-point Likert scales survey was conducted to examine students’ social presence and interpersonal attraction in the chatbot (adopted from Li et al., 2016).</td>
</tr>
</tbody>
</table>

4. Results and Discussion

4.1 Students’ Behavioral Engagement

There were forty-one students who participated in the online listening practice every day using the chatbot. As previously mentioned, students’ behavioral engagement was measured by the completion rate of the required task. The descriptive statistics of students’ behavioral engagement are depicted in Figure 2. The results revealed that the daily completion rate among students was relatively high and remained stable over the ten days ($Mean = 82\%, Min = 71\%, Max = 88\%)$.

Three key factors have been identified in students’ interviews as the main drivers of students’ high levels of behavioral engagement. First, all 12 students mentioned that the immediate feedback given by the chatbot encouraged them to complete the listening tasks, helping them to reflect and revise their previous inputs during the learning process. Besides, the chatbot alleviated the pressure students might feel while doing the assignments compared
with receiving feedback from a human instructor. For example, Student K mentioned that if they misunderstood the question or gave a wrong answer in front of their instructor, they may feel “anxious and shame”. When students experienced the chatbot providing them with immediate targeted feedback and multiple chances to answer a question, they showed a preference for “interacting with a chatbot while completing tasks” (Student C). Second, students were engaged by the task completion function. After completing the listening tasks, they were able to click a “completed” button and check their task completion status in the chatbot system. Clicking on this button every day provided students with “a sense of accomplishment and motivates students to continue by completing the tasks” (Student B). Clear task instructions from the chatbot helped students understand the learning goals in the daily tasks and “navigate their online learning easily” (Student D), which gave them a sense of mastery over the learning process.

These features were helpful in supporting students’ competence in performing the learning activities, which in turn facilitated students’ continued participation in the online listening learning. This finding is consistent with Niemiec and Ryan (2009) who indicated that students are more engaged in learning when their competence is supported.

**Figure 2.** Student Task-completion Rate.  
**Figure 3.** Journal Completion Rate.

4.2 Students’ Cognitive Engagement

We analyzed the learning journal completion rates to examine student cognitive engagement. Figure 3 shows a downward trend over ten days, with the rate dropping from 29% to 15% on the third day and then remaining low. There were 71% students not participating in writing the learning journal from the beginning.

Interestingly, despite the low level of cognitive engagement in the objective data, students in the interviews commented positively on the learning journal activities. Students valued the learning journal in the chatbot because of the increased sense of interaction in a community and the insights they could gain from others’ experiences. For example, when students recorded their learning experiences, the chatbot quoted from previous students' learning journals, such as “Your seniors last semester also mentioned ‘Today's task was particularly challenging.’” In this case, the learning journal activity fostered a sense of connection with previous students that went beyond just interacting with the chatbot. They also indicated that sharing others’ learning experiences in the chatbot “served as a helpful reminder of important subject knowledge” (Student A) and provided “valuable insights into English listening learning” (Student K).

One explanation for this contradiction between the quantitative and qualitative data could be that students are not aware of the purpose of writing learning journals and its benefits. It should be noted that only 29% of the students wrote learning journals at the very beginning. The low participation indicates that a large portion of students (71%) may not have known that this activity existed. Nonetheless, more than half of the students who initially participated in this activity continued to write learning journals over time. The students who wrote the learning journals appreciated that the chatbot fostered a sense of a learning community by showing the experiences of previous students. Future designs should include an introductory session to the chatbot features and closer collaboration between teachers and the chatbot, with teachers explaining the pedagogical benefits of the chatbot learning activity.
4.3 Students' Emotional Engagement

A total of 33 students completed the surveys. Cronbach's alpha results for the social presence and interpersonal attraction scales were 0.96 and 0.95, respectively. Mean scores for social presence ($M = 3.72$, $SD = 1.02$) and interpersonal attraction ($M = 3.61$, $SD = 1.06$) were relatively high (see Table 2), indicating students perceived the chatbot as intelligent and interacted effectively. However, the third item of interpersonal attraction had the lowest mean of 3.48 ($SD = 1.06$), suggesting students held neutral attitudes towards spending additional time conversing with the chatbot after task completion.

Table 2. Social Presence and Interpersonal Attraction Scales

| Scales                  | Items                                                                 | Mean (SD)  
|-------------------------|----------------------------------------------------------------------|------------
| Social presence (n = 33)| 1. I felt as if I were interacting with an intelligent being.        | 3.76 (1.17)
|                         | 2. I felt as if I were accompanied by an intelligent being.          | 3.76 (1.06)
|                         | 3. I felt that the chatbot was able to respond to me effectively.    | 3.67 (1.05)
|                         | 4. I felt involved with the chatbot.                                | 3.67 (0.96)
|                         | 5. I felt that I was able to communicate effectively with the chatbot.| 3.73 (0.94)
| Interpersonal attraction(n = 33)| 1. I liked the chatbot.                                          | 3.70 (1.07)
|                         | 2. I think I could work with the chatbot.                          | 3.76 (0.97)
|                         | 3. I would like to spend more time with the chatbot.                | 3.48 (1.06)
|                         | 4. I think the chatbot could be a friend of mine.                   | 3.52 (1.15)

The interviews with students revealed several factors that were emotionally attractive and aversive. Students enjoyed interacting with the chatbot because the “conversations with the chatbot were relaxed and fun” (Student H) and the chatbot provided “a sense of companionship” (Student E). Two design elements contributed to their emotional engagement: Emojis and the storage of prior learning records. Emojis made the chatbot lively, while reviewing learning records provided students a sense of coherence in their learning journey.

A major shortcoming, however, was the chatbot's repetitive responses. All 12 students reported the pre-set repetitive responses that “made the difference between interacting with the chatbot and a real human” (Student F). The chatbot used the same prompts daily to guide students through the tasks, potentially diminishing their willingness to invest additional time interacting with the chatbot.

For future designs of using chatbots to facilitate students’ emotional engagement, teachers are advised to employ a variety of prompts for the learning tasks and avoid repetitive instructions. The topics of the interaction can be beyond subject knowledge. For instance, chatbots can start the conversation with casual conversation rather than directly introducing the learning tasks. Non-verbal cues (emojis) can be used to express the chatbot's feelings.

5. Limitation and Conclusion

The current study has several limitations that suggest opportunities for future research. First, the short study duration of ten days could be extended in future research to examine student engagement in chatbot-supported learning over the long term. Second, the small sample of 41 students may not be representative of other contexts. Therefore, future studies should include larger and more diverse samples to evaluate the effectiveness of chatbots in online language learning.

In the present study, we designed a chatbot grounded on the SDT components for students’ online English listening learning. Overall, students hold a positive attitude toward the implementation of the chatbot, which was suggested by their behavioral engagement, cognitive engagement and emotional engagement. However, the chatbot is not yet mature enough. Students proposed valuable insights in terms of both pedagogical and technological aspects, which will further help us to develop chatbots in the next round.
Acknowledgement

This research was supported by the Research Grants Council of Hong Kong RGC Research Fellow Scheme (Reference no: RFS2223-7H02).

References


Impact of Self-analysis Behaviors in GOAL for Japanese High School EFL Learners

Zixu WANG*, Rwitajit MAJUMDAR b & Hiroaki OGATA b
aGraduate School of Informatics, Kyoto University, Japan
bAcademic Center for Computing and Media Studies, Kyoto University, Japan
*wangzixu1999@icloud.com, majumdar.rwitajit.4a@kyoto-u.ac.jp, hiroaki.ogata@gmail.com

Abstract: In the 21st century, self-directed learning (SDL) has become an essential skill for personal growth and success. This applies to the context of language learning as well. To enhance students’ SDL skills, we previously proposed the DAPER (Data, Analyze, Plan, Execution, Monitoring, Reflect) model and developed the GOAL (Goal-oriented Active Learner) system to support DAPER. In this study, we focused on the Analysis phase of GOAL. 113 students tracked their scores in 39 weekly tests (19 for English vocabulary and 20 for English grammar) over one year in GOAL. Their self-analysis behaviors after studying and taking tests were also prompted in GOAL. We investigated the relationship between analysis behavior and test scores, finding a significant difference in learning outcomes between students who conducted self-analysis and those who did not. While the current result cannot definitively conclude that self-analysis improves study performance in language learning, it does show that it is meaningful to promote students’ SDL behavior tends to perform better.

Keywords: English as Foreign Language (EFL), Self-analysis, DAPER, GOAL, Self-directed learning

1. Introduction and Motivation

Self-directed Learning (SDL) is an essential skill in the 21st century (P21-Framework, 2009). Significant advances have been made in the methods we acquire knowledge, and there is an explosion in the volume of resources. In this context, promoting SDL skills with technology support has become crucial (Teo, 2010). Analysis skills, an essential part of SDL, can help learners understand their learning progress better, including study habits, progress, outcomes, and evaluating the efficacy of their study strategies (Noguchi & McCarthy, 2010).

Many programs focusing on developing students’ SDL habits often discuss teaching methods and strategies, overlooking other vital factors in SDL like personal needs, self-planning, and self-evaluation (Mia Victor, Walter Lockhart, 1995). While as a component of SDL, self-assessment has been shown to help students improve their speaking skills in English (Chen, 2008), self-analysis exploring personal behavior patterns of achievement has not been studied for its effects on language learning. To fill this gap, we analyzed the impact of self-analysis on the outcomes of English vocabulary and grammar learning, investigating the following research question: What is the relationship between students’ self-analysis habits and their performance in English vocabulary and grammar learning?

2. Goal system for supporting weekly language tests and preparation

The Goal Oriented Active Learner (GOAL) system (Majumdar et al. 2018), is a student-oriented digital platform that focuses on enhancing students’ Self-directed Skills (SDS). Using the GOAL system, students can track their activities, analyze activity data, set goals,
execute monitoring, and reflect on a wide range of activities related to their studies and health, encouraging them to take ownership of their learning processes. The GOAL system can record and visualize students’ past activity data, allowing them to confirm their progress and easily grasp their situation.

2.1 Self-analysis in goal

As an essential step in DAPER, users can conduct simple analysis tasks on the collected data related to an activity. Through this, users can understand their behavior and identify individual problems. The GOAL system offers data input form (for weekly test scores) or synchronizes data from online platforms (for eBooks reading activity) for the data collection phase. For the analysis phase, data is visualized to support users to analyze more clearly and simply. Broadly, the system compares current data with past data or with group averages.

In language learning, the GOAL system offers three parameters (average time, total time, and digitalized learning outcomes) to evaluate the test time score and study time. Scores are evaluated by comparing current test results with past results, and class averages. Students can confirm these data, input comments, and express expectations for performance in the next test.

2.2 GOAL interface

In the language learning part of the GOAL system, students can confirm their SDL skill levels and information related to their scores and preparation time in past tests, as shown in Figure 1.

![GOAL interface screenshot](image)

*Figure 1. The students’ workflow in the GOAL system for test score registration*

The students can input their weekly test scores and the time they spent preparing for a particular test. Then they can check the average test score of their class and get feedback from the system. They can analyze the score and time with the system prompts as shown in Figure 2.
2.3 GOAL learner model

In this study, the goal learner model records all student data from each vocabulary and
grammar test, including the student's UUID (Universally Unique Identifier), a serial number
automatically assigned within the system, preparation time for the test, test scores, feedback
on test difficulty, analysis of test scores, and hopes for the next test. However, this data is
only recorded if the student completed these activities. These activities are not mandatory;
when a student is absent for a test, their score is recorded as null.

3. Research context and findings

3.1 Study context and analysis method

This research was conducted in a Japanese high school. Students were given English
vocabulary and grammar learning tasks independent of the regular classroom curriculum.
The material for learning English vocabulary was a word book called LEAP (Takeoka, 2018),
and the material for learning grammar was called Vintage (Shinoda & Yoneyama, 2018).
Vocabulary and grammar tests were conducted every week. Each vocabulary and grammar
test consisted of approximately 50 questions, and the total score was 20 points. Students
needed to study independently to achieve better results in university entrance examinations.
Teachers introduced the GOAL system and its usage at the beginning of the term. However,
the use of GOAL was not compulsory but just advised to keep track and analyze students’
scores and efforts concerning time spent preparing for the test.

We collected data from 2021 for weekly English tests for 113 second-year high
school students in Japan. These tests included 19 vocabulary tests and 20 grammar tests.
Each test had its scope determined beforehand, allowing students to prepare for the test in
advance. After eliminating invalid data (such as duplicate inputs and test data from the
administrator), we obtained a total of 3129 English test score entries from high school
students and 751 logs related to students’ analysis of their exam scores. Of course, some
students may have studied for tests but didn’t record and manage their study time data using
GOAL.
Table 1. Distribution of the collected data and learners in GOAL system

<table>
<thead>
<tr>
<th></th>
<th>logs (learners)</th>
<th>self-analysis logs (learners)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vocabulary test score</td>
<td>1552 (113)</td>
<td>186 (31)</td>
</tr>
<tr>
<td>grammar test score</td>
<td>1577 (113)</td>
<td>204 (25)</td>
</tr>
</tbody>
</table>

From Table 1, compared to the collected activity data, much fewer data were analyzed for these activities by the students themselves. We observed that only a small proportion of students analyzed the data after preparing for the test and taking the test. Figure 3 is a frequency distribution chart where the vertical axis represents the number of people who took the activity that many times in four activities, and the horizontal axis represents the number of times an activity was taken in four activities.

![Figure 3 The frequency of students’ analysis activities for test score](image)

Based on Figure 3.1, we divided into 2 groups: students who have analyzed their scores after both the vocabulary and grammar tests more than 10 times, and another being students who have done less than 10 times. If a student has analyzed their scores more than 10 times in the grammar tests out of a total of 20 tests and more than 10 times in the vocabulary tests out of a total of 19 tests, we can say they have developed a habit of self-analysis. To analyze whether a student who has developed a habit of self-analysis will influence their study performance in language learning or not, we calculated the average of the scores of Group A, which has analyzed more than 10 times, and Group B, which with fewer than 10 times on the vocabulary and grammar tests every time and conducted a t-test.

3.2 The difference in the performance of students based on self-analysis behavior.

The difference in test scores of students with and without self-analysis of test scores after the same tests are as follows.

![Figure 4. The difference in test scores of students with and without the self-analysis](image)


\begin{table}[h]
\centering
\caption{The information of scores of the students with and without the habit of analysis}
\begin{tabular}{|l|c|c|c|c|}
\hline
                      & With self-analysis habit (N= 9) & Without self-analysis habit (N= 104) & Statistics & p-value \\
\hline
Vocabulary Score     & 15.6 (1.92)                      & 12.5 (3.29)                      & 2.76       & 0.003*     \\
Grammar Score        & 17.6 (1.18)                      & 14.5 (6.97)                      & 1.38       & 0.086      \\
\hline
\end{tabular}
\end{table}

4. Discussion and Conclusion

In this study, we analyzed the impact of self-analysis on the outcomes of English vocabulary and grammar learning and found a significant impact in English vocabulary learning through the results. The performance of students with self-analysis habits is higher than those without self-analysis in vocabulary learning. So we can say that having a habit of self-analysis will help students learn English vocabulary better. On the other hand, even if the average of students with self-analysis habits on grammar tests is higher than those without, by the result of the t-test, it cannot be definitively said that there is a difference. And We think that the lack of significant impact on grammar learning might be due to the fact that grammar learning is possibly a more complex process involving the understanding of rules and structures. Vocabulary and grammar may involve different cognitive and learning processes. Self-analysis might be more suitable for vocabulary learning, which requires memory and recall, rather than rule-based learning in grammar. It could also be possible that the methods used to assess grammar skills are not sensitive enough to detect the impact of self-analysis. These reasons have given us direction and insights for future research.

The findings of this study align with the proposal of the DAPER model, emphasizing the importance of self-analysis in the learning process. By analyzing their own data, students can set specific goals, plan their learning strategies, execute them, monitor their progress, and reflect on their outcomes. The GOAL system supports this process by providing a platform for students to record and reflect on their self-directed learning activities. The significant impact observed in vocabulary learning through self-analysis demonstrates the effectiveness of this approach, while the inconclusive results in grammar learning highlight areas for further investigation.

After demonstrating that self-analysis aids students in learning vocabulary and has some impact on learning grammar, we focus on developing the habit of self-analysis in students. Then, we discovered an interesting phenomenon. If we consider that students analyzed more than 10 times have a habit of self-analysis, then those analyzed from 1 to 10 times in each analysis activity are students who are trying to cultivate their habits of self-analysis. We had a hypothesis that these students should have three stages in their self-analysis: not conducting analysis, attempting analysis, and frequently conducting analysis, and then transforming to the next stage as time changes. However, the analysis data from the activities of students shows that they tend to conduct analysis activities during a specific period. This includes both the study time and test scores for vocabulary and grammar tests. But they can't stick with it. The special period here does not refer to a fixed time. each student has a special period, but it's also possible that they don't. After maintaining a certain period of self-analysis activities, they will stop the analysis activity.

The practical implications of this study offer insights into self-analysis that can enhance vocabulary acquisition, encouraging students to actively engage in the process. Simultaneously, the application of the GOAL system reinforces this process, fostering the development of students’ self-directed learning and self-analysis skills. The findings also highlight the need for further research to understand why students may struggle to maintain self-analysis habits, leading to potential interventions to support sustained engagement.
For future work, we aim to analyze why self-analysis shows a higher impact on helping students in vocabulary learning than grammar learning. We also want to verify further self-analysis’s impact on language learning outcomes and others’ learning outcomes. At the same time, we have also collected data on the amount of time students spent preparing for exams, as well as their analysis and reflections on their preparation time. We hope to analyze whether there is a relationship between this and their study performance. Additionally, we want to analyze the reasons why students cannot stick with analysis activities, and based on the reasons, propose methods to encourage students to develop in self-analysis, we aim to provide more suitable support to promote students’ development of self-analysis and SDL skills.

Acknowledgements

This work was supported in part by the following grants: JSPS JP22H03902 and NEDO JPNP20006 and JPNP18013.

References

Pao-Nan Chou， Wei-Fan Chen (2008)： Exploratory Study of the Relationship between Self-Directed Learning and Academic Performance in a Web-Based Learning Environment. Online Journal of Distance Learning Administration, Volume X1, Number I, Spring 2008
Yuan Yuan Yang, Huiyong Li, Rwitajit Majumdar, Hiroaki Ogata (2023): GOAL system for online self-direction practice: exploring students’ behavioral patterns and the impact on academic achievement in the high school EFL context. Journal of Computers in Education.
AI-powered Collaborative Activities for Chinese Vocabulary Learning

Xinyu GUO, Yun WEN*
National Institute of Education, Nanyang Technological University, Singapore
* yun.wen@nie.edu.sg

Abstract: In recent years, Artificial Intelligence (AI) has significantly increased in digital second language (L2) learning, particularly in supporting vocabulary acquisition. However, research on how AI might facilitate collaborative vocabulary learning is still in its new stage. This study works on investigating the effectiveness of AI-recommended contexts in fostering collaborative language learning among young learners. The research employed a self-developed AI-empowered Chinese vocabulary learning system called ARCHe, which was implemented in primary schools in Singapore. A mixed-methods case study approach was conducted with the 2nd-grade students who spoke English as their first language. Preliminary findings indicate that learning Chinese with ARCHe effectively enhances the academic performance of young learners, with the AI-empowered self-generated contexts feature exhibiting a positive impact on collaborative language learning performance. The study offers insights into the integration of AI in digital language learning, with the potential to enhance L2 learning outcomes for young learners.

Keywords: Artificial intelligence, Collaborative learning, Vocabulary learning, Learner-generated Context

1. Introduction

In recent years, there has been an unprecedented increase in the use of Artificial Intelligence (AI) in digital language learning, with the pandemic further accelerating its adoption. AI technologies such as machine learning, automatic speech recognition, and natural language processing (NLP) have been utilized to provide automated feedback on student writing (Wu, Wu & Zhang, 2021; Godwin-Jones, 2022) and personalized reading recommendations (Hsu, Hwang, & Chang, 2013; Xiao & Hu, 2019). In the context of Chinese as an L2, AI has shown significant potential in supporting vocabulary learning through character pronunciation, recognition, and writing (Kuo et al., 2022). While the importance of learning contexts in vocabulary acquisition is widely recognized (Atkinson, 2002; Wen, 2021), AI may facilitate interactions among students, recommends learning resources, and automatically generates contexts to foster collaborative learning. However, there is limited research on investigating the role of AI in enriching contextual information, promoting interactions, so as to enhancing collaborative learning for L2 learners.

An AI-enhanced Chinese vocabulary learning system, ARCHe, was designed and developed by our research team. In the ARCHe-supported collaborative learning activity, group students are instructed to construct sentences using provided scenarios. The system can automatically recommend students’ artifacts, by which group students will be triggered to create, review, and improve their artifacts continuously. The purpose of this paper is to investigate whether and how AI-recommended contexts can help to promote young learners’ collaborative Chinese vocabulary learning. The research questions are as follows:

1. Did the use of ARCHe help to improve young learners’ Chinese character learning?
2. How did AI-recommended contexts promote students’ collaborative vocabulary learning?

2. AI-powered vocabulary learning
Vocabulary learning, as one of the fundamental language skills, has received considerable attention in recent studies on technology-enhanced language learning (Burston, 2015). The advent of mobile devices has provided L2 learners with real-time, convenient, and context-rich learning opportunities, particularly in the realm of vocabulary acquisition. Notably, AI tools have emerged as effective means to enhance vocabulary learning. These tools empower students to streamline their language learning journey by delegating specific tasks and receiving personalized learning experiences tailored to their individual needs and progress (Roxana & Fabián, 2023). Considering the potential of AI language generation systems for vocabulary practice and providing diverse assistance (Woo & Choi, 2021), we hypothesize that incorporating this feature into L2 learning tools can promote collaborative interaction.

3. ARCHe Design

3.1 ARCHe collaborative learning activities

The ARCHe system comprises two components: a home-based pre-collaborative activity (asynchronous) and a class-based collaborative activity (synchronous) (Wen, 2022). In the home-based activity, students use the ARCHe learning system to complete assigned foundation and creation tasks. The class-based activity involves scenario-based tagging and artifact construction as a group. During the class-based activity, students work in small groups, labeling points of interest in scenario pictures and generating group artifacts. The system’s automatic recommendation feature enriches contexts with self-generated artifacts (Figure 1). As students submit group artifacts, the AI-powered system extracts keywords and recommends related learner-generated artifacts for reference, fostering discussions, peer learning, and continuous refinement and improvement of group artifacts.

![Figure 1. Self-generated contexts recommendation enabled by ARCHe.](image)

4. Methodology

This research employed a mixed methods approach with the objective of examining and contextualizing the effectiveness of AI-powered collaborative activities for Chinese vocabulary learning by enriching self-generated-contexts.

4.1 Participants

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Class</th>
<th>Number of students</th>
<th>Number of groups</th>
<th>Target groups</th>
<th>Ability level</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>C1</td>
<td>20</td>
<td>7</td>
<td>T1G1, T1G4</td>
<td>Lower, Higher</td>
</tr>
<tr>
<td>T2</td>
<td>C2</td>
<td>17</td>
<td>6</td>
<td>T2G1, T2G5</td>
<td>Higher, Lower</td>
</tr>
</tbody>
</table>
The study was conducted in two 2nd-grade classes at School A, with a total of 37 student participants. Among them, we selected two groups from each class as the target groups.

4.2 Intervention Procedure
Each class participated in the project for one semester, engaging in class-based collaborative learning activities conducted once every two weeks. Before each class-based session, students were required to complete specific home-based learning tasks. During these class-based collaborative sessions, teachers summarized the performance of the home-based tasks and initially assigned group tasks. With the assistance of the teachers, each group completed a scenario-based tagging and sentence-making task, subsequently reviewed the work of other groups, and provided comments and ratings. The teacher circulated as a facilitator and collaborator throughout the learning process.

4.3 Data Source and Analysis
In this mixed methods study, quantitative data included pre- and post-tests, while qualitative data was collected through various means. Screen recordings, accompanied by audio feeds of student discussions, were captured on mobile devices within each group. Additionally, in-depth structured post-interviews were conducted with teachers and focus groups to provide qualitative insights for in-depth analysis and to interpret the quantitative data. The focus group interviews with young students covered three main areas: group learning task completion, learning mode preferences, and collaboration satisfaction with ARCHe. Moreover, post-interviews were conducted with teachers to gather feedback on teaching using ARCHe and to capture their overall user experiences.

5. Findings

5.1 The use of ARCHe helped to improve learners’ Chinese character learning
The comparison of the pre & post-tests demonstrates that ARCHe benefits student learning. A total of 37 students from School A. Class 1 and Class 2 completed all the pre-and post-tests. In Class 1, the mean pre-test score was 11.85 (SD=3.45, n=20) out of a total of 30 points. Following the intervention, the mean post-test scores significantly increased to 16.15 (SD=6.09, n=20). The paired t-test revealed a significant improvement in performance among Class 1 students (t=3.91, p<.001). Similarly, in Class 2, the mean post-test score was 16.118 (SD=5.86, n=17), which was higher than the mean pre-test score of 12.82 (SD=4.23). The t-test also indicated a significant difference between the two scores (t=3.18, p < 0.05).

<table>
<thead>
<tr>
<th>Class</th>
<th>Test</th>
<th>No. of students</th>
<th>Estimated Marginal Means</th>
<th>SD</th>
<th>t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-test</td>
<td>20</td>
<td>11.850</td>
<td>3.453</td>
<td>3.857</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td></td>
<td>Post-test</td>
<td>16.150</td>
<td></td>
<td>6.089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 2</td>
<td>Pre-test</td>
<td>17</td>
<td>12.824</td>
<td>4.231</td>
<td>3.182</td>
<td>.003*</td>
</tr>
<tr>
<td></td>
<td>Post-test</td>
<td>16.118</td>
<td></td>
<td>5.862</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the group artifacts scores provide evidence that ARCHe’s learning activities effectively improved the quality of group artifacts. As shown in Table 2, both Class 1 and Class 2 students showed progress in creating artifacts throughout the intervention process. While comparing language skills, we found that students made more significant progress in the content of artifacts.

5.2 The recommendation enabled by ARCHe promoted self-context generation
To explore how AI-recommended contexts promote students’ collaborative vocabulary learning, we examined how group interaction took place with the AI-recommended context. The results revealed that the AI-recommendation effectively motivated young students to...
review and learn from their peers' artifacts, so as to promote self-context generation in the collaborative learning activity.

Under the guidance of the teachers, intragroup members began exploring peers' artifacts through the self-generated context recommendation, which contained keywords relevant to their own artifacts. During the first lesson in intervention, Teacher 1 and Teacher 2 as collaborators, actively encouraged every group to utilize the self-generated context recommendation. For example, after group 6 of class 1 completed generating the artifacts "我有一个姐姐 (I have a sister)," Teacher 1 suggested: "Let's look at other sentences about '姐姐 (sister).'" Then all group members agreed with this suggestion, reviewed and read out one sentence from the self-generated contexts recommendation: "妹妹拿着东西在房间里玩 (my sister is playing with something in her room)" and appraised that was a good sentence.

To facilitate further discussion, we observed that high-level and low-level learning groups employed different strategies in utilizing the AI-recommended contexts. The verbal exchanges from four individual groups are presented in Table 3. For the low-level groups, the recommendation proved more helpful in providing inspiration for generating additional group artifacts. On the other hand, the high-level group, leveraging their judgment skills based on their language proficiency, preferred to comment on and evaluate their peers' artifacts using the recommendation. Their group work through the recommendation primarily focused on improving existing artifacts rather than generating new ones.

Table 3. Examples of intragroup interactions took place in 4 target groups when using the AI-recommended contexts

<table>
<thead>
<tr>
<th>Focus group</th>
<th>Level</th>
<th>Group Talk Content</th>
<th>Analysis</th>
</tr>
</thead>
</table>
| T1G1 | Low | (Click "动物 [Animal]" from self-generated context recommendation)  
- Group member 1: 他用的是 ‘真奇怪’。
  [They used ‘so weird’]  
- Group member 2: 这个动物的句子好像有看过。  
  [I saw this sentence before]  
- Group member 3: 我们可以学这些写。  
  [We can learn from these sentences and write similar one] | All three students of this group were engaged in the talk. They prefer to write more similar sentences by browsing peers' artifacts. |
| T1G4 | High | (Click "雨伞 [Umbrella]" from self-generated context recommendation)  
- Group member 1: 我们没时间了。  
  [We don't have enough time]  
- Group member 2: 让我们再看看有没有更好的句子。  
  [let's see whether others have better sentence.]  
- Group member 1: 我觉得我们的比较好。  
  [I think our sentences are better] | This group reviewed and evaluated peers' artifacts by comparing with their own artifacts, which promote more meaningful discussion during collaboration. |
| T2G1 | Low | (Click "晚上 [Night]" from self-generated context recommendation)  
- Group member 1: 喔，很多很多句子，等一下我们可以 copy 他们的句子。  
  [wow, lots of sentences. Later we can copy their sentences.]  
- Group member 2,3: 好。  
  [OK] | Since this group have limited language ability, they tried to find inspiration of artifacts generation by reviewing peers' artifacts. |
| T2G5 | High | (Click "猴子 [Monkey]", "朋友 [Friend]", "很多 [very]" from self-generated context recommendation) | All students in this group enjoyed |
In the post-interview, regarding the AI-recommended feature, all students expressed their enjoyment of clicking on labels and reviewing their peers' sentences. The two teachers also agreed that this feature stimulated group discussions, idea sharing, and peer learning, enhancing collaborative learning among young learners.

Furthermore, the enthusiasm for utilizing AI-recommended contexts varied between the focus groups from Class 1 and Class 2. Class 1 students displayed curiosity about their peers' artifacts and enjoyed discussing these sentences within their groups. This observation indicates potential differences in teaching approaches between Teacher 1 and Teacher 2. Post-interview findings revealed that Teacher 1 had a consistent focus on developing students' collaborative learning skills from primary 1 up to the current stage. In contrast, Teacher 2 did not extensively discuss collaborative learning during the interview. These findings underscore the importance of considering students' motivation levels and teachers' technological and pedagogical knowledge when integrating AI into language classrooms.

6. Conclusion and limitations

The study demonstrated that the collaborative vocabulary learning activities facilitated by ARCHe effectively promoted young learners’ academic performance and showed progress in creating group artifacts throughout the intervention process. The positive feedback received from both students and teachers indicates that young learners enjoyed engaging in collaborative vocabulary learning with ARCHe, and teachers observed improvements in students' collaborative learning performance.

The utilization of AI-recommended contexts promoted young students to review and learn from their peers' artifacts. It also enhanced collaborative performance by stimulating self-context generation by idea sharing and discussion. The teachers' guidance in encouraging students to utilize the recommendation function played a vital role in this process. Additionally, the variation in enthusiasm observed between the focus groups from different classes highlighted the importance of effective teaching pedagogy and the role of teachers in fostering successful collaborative learning experiences.

Moreover, the study revealed that different strategies were employed by high-level and low-level learning groups in utilizing the recommendation, with the focus being on artifact generation for the low-level groups and artifact improvement for the high-level groups. This indicates that the effectiveness of using automatic recommendation to modify group artifacts depends on the learners' level of competence. Compared to have students reviewing the self-generated contexts recommendation sentences during the collaborative learning process by themselves, the part of teacher's comments and summary of all groups' artifacts were more efficient for helping lower-level students' artifacts improvement.

There are some limitations to this study. Firstly, due to the page constraints, we were unable to provide extensive details about the students’ group learning processes. Secondly, the
effectiveness of students’ character learning was only reported based on students’ pre-and post-test scores. Additional details will be reported in our forthcoming studies.

Acknowledgements
This study is funded by Education Research Funding Programme, National Institute of Education (NIE), Nanyang Technological University, Singapore, with project no. DEV 05/21 WY.

References


Design and Development of a Sentence Construction Game for Deaf and Hard of Hearing (DHH) Users: A Qualitative Usability Study

Arjun PRASAD*, Sunny PRAJAPATI*, Utkarsha, Vishwas BADHE*

Indian Institute of Technology Bombay, India
*22d1204@iitb.ac.in, 22d1201@iitb.ac.in

Abstract: Concerns about the writing proficiency of deaf students in the English language are well known and persistent. Many of them prefer to write concise sentences with fewer adjectives and adverbs and exhibit a variety of syntax usage errors. The purpose of this study was to evaluate the usefulness of a learning intervention created to help DHH students who struggle with sentence formation. The users played with the digital prototype and the perception of the usability of the intervention was analysed and represented using a thematic analysis. Primary insights on the usability and design of this game-based early intervention revealed that the content was well received but more gamification components must be added for the intervention to be both enjoyable and pedagogically valuable. Incorporating more gamification components, conjugating verbs, and studying the effectiveness of the intervention in addressing the difficulty of DHH students in framing sentences will be the focus of the future extension of the study.

Keywords: DHH students, sentence construction, learning intervention, gamification

1. Introduction

The educational experiences of Deaf and Hard of Hearing (DHH) students are significantly diverse and unique compared to students in general education classrooms and students from other disability groups. The Indian educational system has to be strengthened to cater their educational needs as language acquisition remains a major challenge in their life (Mohanty & Mishra, 2020). Studies reported that many DHH individuals struggle with the process of writing and different aspects of writing including phonology, morphology, lexicon, grammar and syntax (Mayer & Trezek, 2019).

Further, DHH children who have limited or no access to spoken language sounds may also find it difficult to learn the rules and patterns of a language like English, particularly because English orthography uses phonemic alphabets (Cannon & Kirby, 2013). Many students also engage in gestures right from a young age to communicate their ideas, which may also contribute to their difficulty in sentence construction during formal education.

Teachers employ a variety of resources to enhance instruction and learning in classrooms, including the use of YouTube videos, guided activities and experience-based learning. Teaching tools and artefacts, such as flashcards, diagrams and rate of speech and hearing charts are used in many classrooms. In contrast to traditional classrooms, it is frequently challenging to teach a full lesson to the students using just speech and gestures. As a result, the key issues facing educators of DHH children continue to be broadening the scope of what may be communicated to the pupils and bridging the gap between what is stated and what is understood (Mohanty & Mishra, 2020). Hence, to meet their educational demands, in addition to better teacher training programs, the teaching materials must be carefully designed. Studies have shown that early intervention helps youngsters develop their linguistic, communication, cognitive, and social abilities more successfully (Marschark et al., 2011).
children who receive early intervention programmes perform better than their peers who do not in terms of language, social development, and early academic achievement (Mohanty & Mishra, 2020).

2. Literature review

A need identification study conducted by Adnyani et al., 2021 highlights the emphasis on the development of multimedia-based English learning materials in addition to highlighting the studies that have put forth the need for incorporating images, visual scaffolds, fingerspelling, etc. to enhance the learning. In another study by Chan et al., 2022, a systematic investigation on French literacy was done for Special Education Needs (SEN) and DHH students with serious games. The study also emphasizes on the improvement in the cognitive skill provided by games to such user groups. One of the major reasons pointed out with respect to poor linguistic comprehension is the lack of syntactic and semantic knowledge by an individual.

The focal question was on how to design an early intervention to improve the level of formal communication in DHH students by addressing their challenge to frame sentences involving conjugating verbs. Studies have recommended the usage of gamification approach in activities involving repetition to keep up the motivation of the user (Faiella and Ricciardi, 2015). Little evidence is available on whether the development of speech or sign language or a combination of both would be beneficial for DHH children. Hence, making an educational programme that defies homogeneity and caters to the unique educational needs of each category (age group/ grade) of students with hearing loss may hold the key to finding the solution. In this paper, we attempt to give an overview of the design and usability of a learning intervention focused on teaching elementary English sentence construction among DHH students which can be further extended to address the challenge involving sentence construction using conjugating verbs.

3. Design and Development of the intervention

The broad objective was to design an intervention that supports special needs education (formal/ informal) of DHH users. The design and development of this game-based learning intervention was carried out in three phases by a group of four researchers as part of the Human Computer Interaction course. The phases included: 1) Needs Analysis, 2) Ideation and Conceptualization, and 3) Prototyping. These are elaborated as follows:

3.1 Needs Analysis

The researchers interacted with DHH students at a school with the help of the teachers and conducted a semi-structured interview with the teachers. Further, three therapists who provide early intervention to DHH students for their learning needs were also interviewed through a semi-structured interview. The discussions were capitulated in the form of a concept map providing us a peek at the interconnections in various aspects of their life and guided us towards identifying a problem, through a 5 Why analysis, that appeared important and addressable given our expertise. The key issues identified include difficulty in constructing formal sentences in any language other than sign language, difficulty in engaging in in-person conversations, difficulty in memorizing long sentences, for example, while learning topics in history as part of school curriculum, where formal long and complex sentences in English are essential. Further exploration through fishbone diagrams and empathy maps was done to get deeper insights into the problem. Additionally, a user persona was also prepared to guide the development of the intervention. Grounded on these analyses, the core problem statement was refined as "Design of an early intervention to improve the level of formal communication in DHH children by addressing their challenges to frame sentences involving conjugating verbs".
3.2 Ideation and Conceptualization

The four group members discussed the concept map and interview data thoroughly and then utilised brainwriting technique to generate 48 ideas. These ideas were then grouped based on the major tools/technologies they utilised. Further development of ideas and the addition of new insights and inspirations also occurred. Following this process, a Decision Matrix analysis was performed to obtain the optimal solution. Each of the four members conducted the analysis individually, based on the collectively decided factors, which were later incorporated into a single Decision Matrix via averaging. The outcome of this process was the finalisation of the idea brief – “Develop puzzles/games (tangible/digital) that teach English grammar and sentence construction”.

Following the ideation process, the paper prototype was created to further understand the interactions associated with the elements of the puzzles. Along with the playability aspect of the game, the paper prototyping also helped in conceptualising the sequence of events, which will address the learning objective. The paper prototyping process was informed by the pedagogical references and the gameplay ideation.

3.3 Prototyping

The prototype of the puzzle was further developed using Scratch, a no-code platform suitable for creating educational tools and games. This allowed us to add many game design elements, such as rewards, levels, goals, rules, and progress. The focus of the team at this stage was to create a functional prototype. Figure 1 shows the Scratch coding environment and the user interface for the game prototype, which was then tested with users. The game’s objective was to help users create sentences using blocks containing words, shaped like puzzle pieces that best describe a given image on the screen. The prototype game had two levels, which is made available on the Scratch project page (Level-1: https://scratch.mit.edu/projects/743654775/fullscreen/; Level -2: https://scratch.mit.edu/projects/755127867/fullscreen/).

4. Gameplay and Scaffolding

In the game, the user was presented with puzzle pieces on which words were written, which were clustered together into three different bins based on the parts of speech they belonged to. In this iteration, these bins were subject, verb, and object. The content in the standard textbooks for grades 1-2 also supported such a presentation. Upon initial instruction, the user is expected to assemble the puzzle pieces to form a sentence of subject-verb-object (SVO) type (see Figure 1).

![Figure 1: Screenshots of the game screen developed on Scratch](image-url)
Upon successful completion of a sentence, feedback is generated by the system (see Figure 1 (4)), and the screen is refreshed for the next sentence activity. The scaffolding in the games was provided using visual elements, such as shapes and colours. Figure 2 indicates the scaffolds at different levels, i.e. for level 1, the scaffolds were provided in terms of shape and colour of the puzzle pieces, so that words corresponding to a specific part of speech will have different colour and shape. These elements will gradually fade away (colour followed by shape) as indicated in Figure 2. A progress bar indicates the remaining number of puzzles to be solved in each level.

![Figure 2: Scaffolds used at different levels of the game](image)

5. Methodology (User testing)

The usability evaluation of the prototype was conducted with five users in a laboratory environment. Out of these five people, two were children (male) between the ages of 6 - 6.5 years without any hearing impairment, whereas the other three belong to the age group of 18 years and above, with severe hearing impairment to complete deafness. The variation in age group of the participants is due to our limited access to young DHH children. Users 1 and 2 helped us understand the usability challenges faced by young children while playing the game. Users 3, 4 and 5 helped us understand the challenges faced by DHH users while playing our game. We acknowledge this as a limitation to be addressed in future studies. Despite this limitation, since this was a preliminary study, data from these users helped us refine our game for young DHH users.

The two users of 6 and 6.5 years of age received instructions on how to play the puzzle in both English and Tamil. User 1 had some initial difficulty in positioning the pieces and navigating the information on the screen, but was able to complete Level 1 successfully. User 2, who had prior knowledge of puzzles, was able to quickly solve the first puzzle and successfully completed both the levels with the help of the progress bar. User 2 was able to navigate between the two levels and figured out the changes in the puzzle shape in Level 2. User 1 showed interest in moving to Level 2, but was not willing to play with the pieces due to the puzzle appearance.

The remaining three users with severe hearing impairment to complete deafness were given instructions to play the game in written format, as indicated in Figure 3. They completed both the game levels with minimal difficulty; post this study their feedback on the game was obtained by communicating with gestures and writing.

![Figure 3: Interaction with DHH people with severe hearing impairment to complete deafness](image)

6. Results and Discussion

Educational games, as visual instruments, play a crucial role in facilitating knowledge acquisition among DHH students and leveraging a strong visual approach in the creation and utilization of materials can greatly enhance the learning journey of such users. According to Bouzid et al., 2015, many of the applications developed for DHH students do not possess essential gameplay elements, such as well-defined objectives and scoring systems, which are necessary for them to be accurately classified as complete games.

The researchers tried to incorporate several gamification elements in this learning intervention and in the present study, the usability of the intervention was examined that
produced several key takeaways regarding the gamification aspects. A thematic analysis (Braun & Clarke, 2006) of the study revealed the main four themes regarding the perception of the usability of the intervention: (1) consistency of images (2) navigation within the game (3) feedback, and (4) question format. The user responses varied for themes (1) and (4) but shared more similar aspects for themes (2) and (3).

Theme 1: Consistency of images:-- Users 1, 4 & 5 were able to map the given image to corresponding words (visual consistency) but user 2 suggested ensuring consistency in the mapping of pictures to different forms of verbs. User 3 expressed the need to increase the image size and use of animated images. User 1 expressed a confused tone while saying ‘Is the boy eating a cake?’ User 2 asked for clarification: ‘Is it Rahul? A person is playing chess. Is it she?’ The results show that more consistency in the images are required to improve the user experience of the intervention.

“Theme 2: Navigation within the game:-- The users could easily navigate between both the levels of the game. Users also found the progress bar and option to skip the questions useful. Simplicity in navigation is crucial in any online game and the results showed users appreciating the basic features incorporated in the intervention.

“| The game is interesting” (User 1) |
| “I am going to level 2. The last one is this. I see the dot.” (User 1) |
| “In both the games, I completed all the dots.” (User 2) |

Theme 3: Feedback:-- Most of the users emphasised the need to improve the reward system, for example, incorporating animated images as suggested by user 3. Also, most users suggested incorporating better feedback techniques. The users also noted that the progress bar was helpful while playing with the intervention. Providing timely and informative feedback is crucial to keep the motivation high in users and contribute towards learning.

“What is this tick? Did I make everything correct? Did I win?” (User 2)

Theme 4: Question format:-- Users 4 & 5 suggested adding a picture and sign for each letter on the puzzle in addition to exploring other formats including fill-in-the-blanks and match-the-following (to make the intervention similar to applications like Duolingo). User 3 observed that the game was appropriate for classes 1-2. Users 4 & 5 suggested that the game was appropriate for classes up to grade 4. Further, user 1 & 2 could identify the difference in scaffolds used in questions in both levels.

“Oh, what to do? You have to arrange it (puzzles in level 2). I don’t want to do it.” (User 1)
“This is box-box now.” (User 2)

7. Conclusion and Future work

The aim of this study was to understand the usability of a learning intervention designed to address the difficulty in sentence construction among the DHH students. The users played with the digital prototype and the results of the study provided primary insights on the usability and design of this game-based early intervention. Though the content delivery of the intervention was found to be appropriate, more elements of gamification including clear and concise instructions, help functions, and control over gaming options such as speed, and difficulty need to be incorporated. This is crucial in bringing in engagement, interactivity, and active participation among the users though there still exists ambiguity in identifying what features would a learning game need to be both entertaining and pedagogically useful. Further, the intervention could be used by any novice learner with a motivation to learn the English language. Future research is proposed to focus on incorporating more features of gamification, and conjugating verbs and study the effectiveness of the intervention in addressing the challenge to frame sentences involving such forms of verbs.

Acknowledgment

We acknowledge the guidance given by Prof. Chandan Dasgupta and contribution of Shirish Chinchanikar, Nisumba Soodhani K and Meera Daulatrao Pawar for their inputs during the
conducted study. We also thank the users who participated in the user-testing on the intervention.

References


Using T-Robot Board Games to Enhance Learning Gains of Rural Elementary School Students

Yu-Wei WU a*, Wen-Yu YE b*, Wen-Chi Vivian WU c* & Rong-Jyue WANG d*

a Department of Foreign Languages and Literature, Asia University, Taiwan
yuwei530@gmail.com

b Department of Foreign Languages and Literature, Asia University, Taiwan
wunyu89@gmail.com

c Department of Foreign Languages and Literature, Asia University, Taiwan
vivwu123@asia.edu.tw

d Department of Electronic Engineering, National Formosa University, Taiwan
orffwang@nfu.edu.tw

Abstract: This study used T-Robot board games to create an interactive and enjoyable English language learning environment for the purposes of reducing learners’ anxiety while also enhancing their speaking skills. The participants of this study were 28 fourth-grade students with limited English proficiency from a rural area of Taichung City, Taiwan, and multiple data sources using mixed methods were also adopted for data analysis, including pre-and post-tests, questionnaires, and interviews. Quantitative results showed significant improvements in both speaking ability and reduced anxiety, and positive qualitative feedback was also displayed from the interviews. The integration of robotic board games provided a promising alternative to traditional teaching methods, particularly for lower-proficiency students.

Keywords: Robot board game, Speaking anxiety, Enjoyment

1. Introduction

In today’s globalized world, English proficiency is crucial for success. However, the rural area of Taichung City faces a challenge with low English proficiency among elementary students, particularly in speaking, leading to heightened anxiety in communication. Ho (2010) highlighted how traditional teaching methods and reliance on commercial materials hinder cognitive development and critical thinking skills. To overcome this, an innovative approach incorporating board games and robots into the curriculum has been explored. Notably, Huang Qianyi’s 2018 study introduced board games tailored for elementary students to enhance English conversation skills. The combination of these games with the engaging T-Robot aimed to improve learning enjoyment and reduce anxiety. By offering an interactive and appealing English learning experience, the goal was to elevate proficiency and foster a greater sense of enjoyment in the learning process. To investigate the effectiveness of the proposed teaching method, the following three research questions guided this study:

1. To what degree did robot board games alleviate elementary school students’ oral communication anxiety in English language learning?
2. To what degree did the integration of Robot board games improve students’ English-speaking proficiency?
3. How did students perceive the use of Robot board games in enhancing their enjoyment of speaking English?
2. Research methods

The study utilized a robot board game as a teaching tool to integrate English language learning for 29 fourth-grade students in the rural area of Taichung City. The research aimed to evaluate the learning effect and feedback of students after a 10-week experiment, consisting of two 40-minute classes per week. In the first week, students were administered a pre-test of English-speaking proficiency and English ability, along with a pre-questionnaire on anxiety. Throughout the teaching process, illustrated in Fig. 1, the teacher began by teaching the necessary vocabulary for the students to apply in their work. Next, students were instructed to work collaboratively in groups to create a school map, incorporating places, treasures, monsters, and a route, while simultaneously practicing the oral speaking part for their final presentation, which would take place during weeks 2-7. During the final week, students completed a post-test of English-speaking proficiency and ability, as well as a post-questionnaire. The study used pre-test, post-test, and questionnaires as assessment tools, including the modified test and questionnaire which was adapted from Luo, K. M. (2019). The speaking test comprised three parts: cloze, multiple choice, and oral tests, and the modified questionnaires included 10 questions items. In addition to quantitative analysis, the study also conducted qualitative one on one interviews with 12 students and classroom observations to gain an in-depth understanding of this innovative instruction. The interview protocol asked two questions about students' enjoyment of the course and their feelings about the robot. The study sought to answer three research questions related to the improvement of English-speaking anxiety, enjoyment, and English-speaking proficiency score.

Fig. 1 Students collaborating as a team to use coding cards and make T-Robot move

2.1 T-Robot (Assisted by Reading & Rhythm Co., Ltd.)

T-Robot, a robotic board game, utilized coding cards for directing the robot's movement. The game involved students interpreting and executing commands from these cards on a map. With dimensions of 106x71x71 mm, the robot was designed for easy handling, featuring wheels for smooth movement. Command cards provided visual symbols and English explanations for user-friendly interaction.

3. Findings

The study found that incorporating the T-Robot board game into English teaching effectively reduced speaking anxiety, improved enjoyment, and enhanced speaking abilities.

3.1 To What degree did robot board games alleviate elementary school students’ oral communication anxiety in English language learning?

The study showed significant reduction in oral communication anxiety using T-Robots. A five-point Likert scale pre-intervention (M = 3.66, SD = 0.744) and post-intervention (M = 4.16, SD = 1.023) demonstrated a -0.500 difference. The 95% confidence interval (-0.851 to -0.149) supports this change. A t-value of -2.923, 27 degrees of freedom, and p-value of 0.007 confirm the anxiety reduction. T-Robots effectively alleviate language-related anxiety, highlighting technology’s role in addressing student challenges in language learning.
3.2 To what degree did the integration of Robot board games improve students' English-speaking proficiency?

In RQ2, Table 2's analysis for the English-Speaking Proficiency Test revealed a significant reduction in oral communication anxiety (mean difference: -12.857, pre-test: M = 31.43, post-test: M = 44.29, SDs provided). The highly significant t-value of -7.586 (27 degrees of freedom, p = 0.000) reinforces this finding. Improved English-speaking proficiency is indicated by increased post-test scores and reduced variability. These results guide educators towards effective strategies for enhancing learners' oral communication skills.

### Table 2

Pre-test and Post-test Results for English Speaking Proficiency Test

<table>
<thead>
<tr>
<th></th>
<th>Mean difference (Pre-Post)</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>31.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post</td>
<td>44.29</td>
<td></td>
<td></td>
<td>.000**</td>
</tr>
</tbody>
</table>

RQ3 interviews revealed students' enjoyment of robot board games due to their engagement and interactivity. They valued the challenges and teamwork, with one student expressing increased enthusiasm for English class. These findings highlight the games' potential to enhance English learning enjoyment, encourage participation, and improve proficiency, particularly for students with lower language skills.

4. Conclusion

Incorporating innovative teaching methods like using games in the English curriculum has the potential to improve learning outcomes, especially for less proficient students. This study’s results supported the effectiveness of the Robot board game method in enhancing English proficiency, reducing anxiety, and increasing learning enjoyment. Therefore, integrating this method can create a more engaging and effective language learning experience, ultimately benefiting students’ academic and professional achievements.

References


PTP
How teachers’ conceptions of student engagement influenced their actual strategy implementation, and student online engagement

Gulipari MAIMAITIa & Khe Foon HEWb*

a, b Faculty of Education, The University of Hong Kong
Hong Kong SAR, China
* kfhew@hku.hk

Abstract: In this study, we explored the interplay between teachers' conceptions of student engagement, their strategy implementation, and student engagement in a videoconferencing-based fully online learning environment. The mixed-method multi-case study involved two postgraduate courses, with data collected via surveys and semi-structured interviews. Findings illustrated diverse conceptions of student engagement among instructors, influencing their teaching strategies and ultimately affecting student engagement. Instructors with a comprehensive, multidimensional understanding of engagement encompassing behavioral, emotional, and cognitive aspects were more likely to employ effective strategies, resulting in improved student engagement.

Keywords: Student engagement, teacher’s conception, teacher’s knowledge, engagement strategy, synchronous online learning

1. Introduction

The COVID-19 pandemic triggered a transition to online learning, utilizing videoconferencing for synchronous learning. However, research points to low student engagement in this setting (Maimaiti et al., 2021). Engagement, a key factor in learning outcomes and positive behaviors (Martin & Bolliger, 2018), is greatly influenced by the instructor (Farrell & Brunton, 2020). Yet, discrepancies often exist between teachers' conceptions and practices (Ndeke & Keraro, 2021). This study investigates how teachers' engagement conceptions influence their strategies and its subsequent effect on student online engagement. It seeks to answer: (1) How do instructors conceptualize student engagement? (2) How do their conceptions affect their implementation of engagement strategies? (3) How does their strategy implementation affect student engagement?

2. Method

This study, using a multi-case approach, quantitative and qualitative methods, explored two postgraduate courses at a Hong Kong university that had moved to online learning due to the pandemic. Participants, selected through convenience sampling, provided data via interviews and a survey. We conducted online interviews with 24 students (12 students from each course). Each interview took about 35 minutes. Pedler et al.’s (2020) engagement strategies framework (as shown in Figure 1) is used to guide question design. This framework offers instructors guidance for fostering behavioral, emotional, and cognitive engagement. Our research focused on these strategies within teaching practices, addressing gaps in previous studies that either ignored engagement's multidimensional aspects (e.g., Nafukho & Chakraborty, 2014) or concentrated on student strategy use (e.g., Redmond et al., 2018). Along with the interview, we also used a self-reported engagement
questionnaire adapted from Skinner et al. (2008) and Rotgans and Schmidt (2011). This 14-item questionnaire, validated with Cronbach’s alpha values over 0.8 (Lo, 2017), gathered data on behavioral, emotional, and cognitive engagement. We received 47 valid responses, with internal consistency of 0.89.

The interview data was analyzed using deductive content analysis with Pedler et al.’s framework as the coding matrix, and the survey data was descriptively analyzed and tested for significant differences between the two courses.

Figure 1. A framework of guidelines for teachers to promote student engagement (Pedler et al., 2020, p. 55)

### 3. Findings

Instructors from two courses held distinct conceptions of student engagement. Course One’s instructor focused on learning experience design and alignment of activities with learning outcomes, viewing engagement as students mentally navigating through the learning process. Conversely, Course Two’s instructors conceptualized engagement more specifically in terms of behavioral, emotional, and cognitive aspects, underlining the importance of active participation, enjoyment, and content understanding.

Course Two implemented all engagement strategies outlined by Pedler et al. (2020), while Course One’s implementation was more selective. Course One’s instructor used a rigid course outline, limiting opportunities for behavioral engagement. Students desired more interactive activities. Similarly, emotional engagement strategies were deficient in Course One due to the instructor’s lack of emphasis on interaction, making students feel distant and disengaged. In contrast, Course Two fostered positive teacher-student relationships by creating interaction opportunities. Cognitive engagement strategies were partially implemented in Course One, with a focus on self-directed learning but limited measures to ensure student understanding and application of the content and promote student interest. Course Two, on the other hand, utilized all cognitive engagement strategies, employing diverse resources and activities to make lessons interesting, interactive, and practical.

Quantitative survey data of a T-test as showed in Table 1, indicated that students in Course Two exhibited significantly higher levels of behavioral, emotional, and cognitive engagement compared to students in Course One, which aligned with the interview findings.

Table 1. T-test result of students’ behavioural, emotional, and cognitive engagement
<table>
<thead>
<tr>
<th>Engagement dimension</th>
<th>Course</th>
<th>N</th>
<th>Mean</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral engagement</td>
<td>Course One</td>
<td>25</td>
<td>3.592</td>
<td>-4.323</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Course Two</td>
<td>22</td>
<td>4.264</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional engagement</td>
<td>Course One</td>
<td>25</td>
<td>3.3920</td>
<td>-4.211</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Course Two</td>
<td>22</td>
<td>4.0364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive engagement</td>
<td>Course One</td>
<td>25</td>
<td>3.0400</td>
<td>-3.832</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Course Two</td>
<td>22</td>
<td>3.6136</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Accessibility

This research investigated how instructors’ engagement conception affect their strategy implementation and consequently, online student engagement. Instructors’ unique conception, perhaps influenced by their research interests and other elements, significantly affected strategy effectiveness. The instructor of Course One focused on behavioral engagement strategies, leading to lower student involvement due to limited interaction and disregard for student interests. However, Course Two’s instructors, who employed a broader engagement conception, utilized diverse strategies, enhancing student engagement. They fostered behavioral, emotional, and cognitive engagement via group activities, personalized communication, and practical application. The study highlights that instructors’ engagement understanding informs their teaching strategies and affects student engagement online. It implies a comprehensive engagement conception can direct effective strategy execution. Future research could delve deeper into factors influencing instructors’ conceptions of student engagement and involve larger participant sample size.

References


Martin, F. & Bolliger, D.U. (2018). Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning 22*(1), 205–222. [https://doi.org/10.24059/olj.v22i1.1092](https://doi.org/10.24059/olj.v22i1.1092)


ICCE 2023 Exploring the Social Media Discourse: the Impact of ChatGPT on Teachers' Roles and Identity

Yuchun ZHONG a, Davy Tsz Kit NG b, Samuel Kai Wah CHU c
aAffiliation A, The University of Hong Kong, Hong Kong, China
bAffiliation B, The University of Hong Kong, Hong Kong, China
cAffiliation C, Hong Kong Metropolitan University, Hong Kong, China
*sunnyzhong33@connect.hku.hk

Abstract: As the use of ChatGPT in education becomes more prevalent, it is important to examine the potential impact on teachers' roles and identity. This study aimed to investigate how ChatGPT may shape teachers' roles and identity by conducting a thematic analysis of contents from WeChat posts that discuss the use of ChatGPT by teachers. The analysis focused on identifying themes related to the impact of ChatGPT on teachers’ perceived self, pedagogical roles, and relationships with students. The findings suggest that ChatGPT may have both positive and negative impacts on teachers' roles and identity. On the one hand, ChatGPT can bring new possibilities to teachers allowing them to focus more on students' qualities and soft skills development and personalise feedback and assessment for students, which could reduce the workload of teachers and enhance their effectiveness. On the other hand, the use of ChatGPT may threaten the professional identity of teachers, leading to feelings of inadequacy or a loss of control over the teaching process. Furthermore, the use of ChatGPT may give rise to ethical issues in teaching and lead to a trust crisis between teachers and students. This study highlights the need for further research on the impact of ChatGPT and provides insights into the implementation of ChatGPT in education.

Keywords: ChatGPT, Artificial intelligence (AI), social media, teacher identity, teaching roles, thematic analysis

1. Introduction

Research on artificial intelligence (AI) is not new but becomes newly and unprecedentedly fashionable due to the emergence of ChatGPT (Generative Pre-trained Transformer), a powerful natural language processing model developed by OpenAI that can generate human-like text based on a given prompt and can be trained as a conversational agent responding to a wide variety of commands from end users (Jalil et al., 2023). In a remarkably short time following its release, ChatGPT has spurred widespread discourses across a diverse range of societal spheres (A.Shaji George et al., 2023; Biswas, 2023; Surameery & Shakor, 2023), among which the educational sector is one area where ChatGPT has the potential to make a significant impact. Many educators have conveyed enthusiasm about the potential for these tools to introduce innovative pedagogical opportunities (Jalil et al., 2023). Others, however, are concerned about the potential of technology to dehumanise the learning experience driven by fear of commercialisation or neo-liberal ideologies wrapped up in technology (Cox, 2021).

With artificial intelligence and robots providing customised instruction to students, teachers’ roles may shift toward overseers who design and select machine-led instruction, monitor student progress, and provide support (Edwards et al., 2018). Teachers play a critical role in shaping the learning experiences of students, and their professional identity is an important aspect of their work; as such, research on teachers’ professional identity formation contributes to our understanding of what it feels like to be a teacher in the face of rapidly advancing technologies and how teachers respond to these changes (Beijaard et al., 2004). From this point of view, it is crucial to understand how teachers consider their identity in the teaching profession. What is found relevant to the profession, especially with the
increasing introduction of artificial intelligence into education, may clash with a teacher’s values and beliefs about what constitutes effective teaching practices (Beijaard et al., 2004).

2. Literature Review

2.1 Teacher Identity

Teacher identity has been variously defined in the literature (Beijaard et al., 2004). Gee (2000) defined teacher identity as a certain “kind of person” or “several different kinds” acting and interacting in a given context (Gee, 2000, p. 99). However, Hoffman-Kipp (2008) believe that teacher identity is not a fixed point that individuals interact with the world, but ‘the intersection of personal, pedagogical, and political participation and reflection within a larger sociopolitical context’ (p. 153) (Beijaard et al., 2004). While there is no consensus regarding what constitutes teacher identity, there is a general acknowledgement that teacher identity plays a key role in decisions teachers make about their teaching practices, the content they teach, the kind of relationships they maintain with their students (Beijaard et al., 2004), as well as ‘where they place their effort, and whether and how they seek out professional development opportunities’ (Hammerness et al., 2005, pp. 383–384).

In this study, teacher identity is understood from a socio-cultural point of view, which holds that teachers do not develop their identity in a vacuum, but rather in a context that brings social and cultural forces to bear upon that development (Holland & Lachicotte, 2007). The concept of teacher identity encompasses three key dimensions, 1) perceived self refers to the personal and subjective sense of self that teachers develop as they navigate their role as educators, which includes a teacher’s beliefs, attitudes, values, and emotions, 2) pedagogical roles related to teaching and learning that involves the knowledge, skills, attitudes, and values that define a particular profession and distinguish it from others, and 3) their relationships with students (Beijaard et al., 2004).

![Figure 1. Conceptualising teacher identity](image)

2.2 The Impact of ChatGPT on Teachers’ Identity

The use of technology in education has raised concerns about its potential impact on teachers’ roles and identity. For example, some studies have suggested that the use of technology in the classroom can enhance teachers’ professional identity by enabling them to be more innovative, creative, and effective in their teaching practices (Donaghue, 2020; Koehler & Mishra, 2009; Zhong et al., 2022). Other studies have highlighted the potential negative impact of technology on teacher identity, such as feelings of inadequacy or a loss of control over the learning process (DiGregorio & Liston, 2018; Hill & Uribe-Florez, 2020).
While ChatGPT is a relatively new technology, there have been several studies exploring its potential impact on teacher identity, with mixed results. ChatGPT can assist teachers in completing their roles in teaching practices. For instance, the AI-powered system can help teachers provide customized feedback to individual students, saving them time and effort (Lund & Wang, 2023). It can be used to create intelligent tutors that provide personalised guidance and feedback to students as they progress with their studies (Kalla & Smith, 2023; Markel et al., 2023). Despite the benefits, Shidiq (2023) indicated that the role of human teachers cannot be replaced by ChatGPT because education requires direct interaction, emotional connection, creativity, and the ability to capture nuances and learning styles of individual students.

Additionally, the pedagogical dimension of teacher identity may be impacted by the potential for ChatGPT to automate certain aspects of grading and assessment (Kalla & Smith, 2023). For instance, a study using questions from a software testing book by Ammann and Offutt indicated that ChatGPT can provide correct or partially correct answers in 55.6% of cases, provide correct or partially correct explanations of answers in 53.0% of cases, and that prompting the tool in a shared question context leads to a marginally higher rate of correct responses (Jalil et al., 2023). However, Tack and Piech (2022) have developed a solution for testing the abilities of an AI teacher by running conversational agents in parallel to human teachers and comparing their responses in terms of speaking, understanding, and helping students. The findings indicated that while AI performs well on conversational uptake, they are quantifiably worse than real teachers on several pedagogical dimensions, especially with regard to helpfulness (Tack & Piech, 2022). Pickell & Doak (2023) have highlighted that while integrating ChatGPT into teaching, instructors should focus on creating better assignments that require critical thinking and analysis rather than a mere recollection of facts and that they should remain vigilant in detecting plagiarism and cheating.

It was also noted that the introduction of ChatGPT may affect the teachers’ relationship with students. ChatGPT can be used to develop innovative projects and resources and create interactive games and activities that engage students more meaningfully (Kalla & Smith, 2023). A study by Markel et al. (2023) found that ChatGPT provides the opportunity for teachers to get valuable teaching practice without the pressures of affecting students, allowing them to iterate their responses both during each session and across sessions. ChatGPT can simulate social dynamics in student-teacher teaching settings, giving teachers the chance to practice mediating peer-peer learning interactions in a safe setting.

2.3 Research Questions

Given the nascent nature of ChatGPT, there is currently a lack of solid empirical studies related to how the use of ChatGPT may shape teachers’ identity. This paper aims to fill this gap by conducting a content analysis of websites to investigate how ChatGPT may shape teachers’ identity. Specifically, the study examined how ChatGPT may impact teachers’ beliefs, attitudes, and values related to their role as educators, and how it may dynamically shape teachers’ “collective” identity. The following research questions guided the inquiry:

RQ 1. In what ways does ChatGPT impact the roles of teachers?
RQ 2. How do teachers integrate ChatGPT into teaching pedagogies?
RQ 3. How does ChatGPT change teachers’ relationships with students?

3. Methodology

3.1 Research Design

In this study, we employed thematic analysis to analyse the contents of posts or articles published on the WeChat platform, a highly influential social media platform in mainland
China. The posts or articles were centred around the potential impact of using ChatGPT on the roles and identity of teachers. Thematic analysis is a qualitative research method that is particularly useful for analyzing the content of textual data (Vaismoradi, et al., 2016). It is a flexible and systematic approach that can be used to provide a comprehensive and nuanced understanding of the underlying meanings and perspectives in the data (Braun & Clarke, 2006).

3.2 Data Collection

To collect data for this study, we conducted a search of relevant posts and articles using a combination of keywords, including "ChatGPT," “teacher," and "education" through one search engine WeChat platform to identify articles that are related to the use of Chat GPT by teachers, rather than crawling all public posts on the use of ChatGPT in education. Specifically, we used the following selection criteria to manually identify websites for inclusion in the study: 1) posts and articles discuss the use of ChatGPT in education and its potential impact on teachers’ identity, 2) posts and articles published by reputable sources, such as academic institutions, educational organisations, or educational bloggers, 3) posts and articles that discuss the use of ChatGPT in education, and that provide information on the potential impact of this technology on teachers’ identity, 4) sources published within two months from February 1, 2023 to March 31, 2023, in order to ensure that the data is up-to-date, and 5) the articles published on the platform have more than 100,000 reads to ensure that the articles selected for analysis have achieved a certain level of popularity or readership on the platform and are representative of the most widely read and influential articles on the platform. As a result, 133 articles and posts were retrieved. We further excluded articles that are identical and too short. Identical articles may indicate a lack of diversity or originality in the content, while very short articles may not provide enough context or details to fully understand the topic at hand. The WeChat platform system requires original articles to have a word count greater than 300. Regarding the nature of the platform, articles that did not meet the word requirement were excluded. Finally, 32 articles met the inclusion criteria after eliminating duplicates. At the time of the final analysis, six sites were defunct, leaving 26 articles and posts in the final sample.

3.3 Data Analysis

In this study, thematic analysis along with a deductive approach was used for analysing the contents of the articles. The deductive approach is a top-down technique that involves starting with a clear framework and using it to guide the analysis of the data (Braun & Clarke, 2012). This study was guided by the conceptual framework on teacher identity (Figure 1), which provides an understanding on the impact of ChatGPT on three dimensions of teacher identity, including perceived roles as a teacher, teachers’ pedagogical identity, as well as teachers’ relationships with students. As shown in Appendix 1, the articles included in the analysis covered one or more dimensions of teacher identity.

The first and second authors of the study examined ideas that are relevant to the framework and then identified evidence of these ideas from the data. We began by conducting a preliminary analysis of the data to identify common themes and patterns under each dimension of teacher identity. We then developed a coding scheme and use it to systematically code the contents (Appendix 2). The coding scheme was designed to capture data related to how ChatGPT may impact teachers’ roles and identity from three dimensions (i.e., changes in teachers’ roles, changes in pedagogy, and changes in the teacher-student relationship). We analysed the coded data to identify patterns related to how ChatGPT may impact teachers’ identity. To ensure the validity and reliability of the study, the two authors used a systematic and transparent approach to data analysis. The authors maintained a detailed audit trail of the entire research process, including decisions made and changes implemented. The data analysis involved several rounds. In the first round, the first author and second author read the articles to gain a comprehensive understanding of themes on
4. Results

RQ 1. How may ChatGPT impact the roles of teachers?

The qualitative analysis of relevant posts on the WeChat platform revealed several subthemes related to the impact of ChatGPT on teachers’ perceived self as a teacher. One of the major concerns was related to whether ChatGPT would replace the role of teacher. Many believed that the development of ChatGPT will pose a threat to the teaching occupation (Source 7). "Due to the ability of ChatGPT to automatically perform repetitive tasks such as grading, providing feedback, and creating educational content, it may lead to some teachers becoming unemployed" (Source 2). Others pointed out that while teachers can not be replaced by AI for the time being due to the difficulty for AI technology to replace human decision-making in complex situations, especially those that require the integration of emotions, experience, and scientific knowledge (Source 4, 6, 23), educators, however, should begin to consider changing their teaching methods when AI is providing an increasingly personalised education for students. As AI will continue to evolve, teachers are compelled to think of their 'irreplaceable value' in the future (Source 2, 4).

Additionally, teachers' role as traditional knowledge transmitters was questioned. "In the traditional teaching model, students mainly acquire knowledge through textbooks and teacher lectures, with limited learning materials and inadequate connection between learning activities and real-life situations. However, in the future of education, students can obtain a more diverse range of knowledge through artificial intelligence technologies such as ChatGPT" (Source 4). The solid mastery of factual knowledge through rote memorization is considered one of the advantages of East Asian education. However, with the emergence of ChatGPT, these advantages turned into disadvantages (Source 1, 4, 21, 22, 24, 25, 26). As a result, "teachers should go beyond knowledge transmission and pay greater attention to the quality of learning" (Source 25). Firstly, "teachers need to explore and transform the educational value of human-machine dialogue, create more diverse learning methods, rethink what to teach and what to learn, and ultimately use information technology to comprehensively improve the learning of students" (Source 24). Secondly, teachers should emphasize more on quality cultivation and soft skills development (Source 20, 22). As it was indicated in one article, "what truly leads to a person's success is the ability to focus on learning, to elevate from interests to aspirations, and to become a unique individual with distinctive personal qualities. It is also reliant on responsibility and resilience, as well as the ability to persevere and persist in the face of difficulties. Additionally, it is dependent on good interpersonal relationships and a spirit of collaboration" (Source 22).

Furthermore, the relationship of teachers with AI technologies was also the focus of many articles and posts. One of the most discussed topics is in what ways will ChatGPT help reconstruct the relationship between teachers and machines. Some viewed it as "an opportunity to reevaluate the value of human teachers and reshape teacher capabilities" (Source 21). Others believed that "in the process of teaching, human teachers and intelligent tools should have a collaborative relationship, namely human-machine collaboration. Human teachers and intelligent assistants should divide their work reasonably, cooperate, and jointly complete teaching tasks" (Source 7).

RQ 2. How would teachers integrate ChatGPT into teaching pedagogies?

As identified in the selected articles and posts, there were several ways of integrating ChatGPT into teaching pedagogies. First, teachers can use ChatGPT to promote
personalised learning. "ChatGPT can help teachers better understand students' learning situations, provide targeted guidance according to their needs, and improve teaching quality" (Source 4, 24). "ChatGPT can understand and generate natural language, possess strong conversational interaction capabilities, and provide students with high-quality and personalised learning content and services, greatly transforming the learning process and methods. For example, ChatGPT can analyse students' learning interests and needs based on their language input and behaviour, generate teaching resources and strategies that match their personalised learning styles, make learning more tailored to students, and improve students' learning enthusiasm and initiative" (Source 4). Second, teachers can integrate ChatGPT with existing tools to promote more interactive teaching. As the functions of ChatGPT continue to evolve, teachers can utilize ChatGPT to generate a wide range of learning resources, such as online videos, games, virtual reality, and other forms, thereby increasing learning interest and motivation (Source 4, 7). Third, the way of providing feedback to students would be more diversified and personalised (Source 4, 10, 26). Teachers can use generative and personalised feedback. "ChatGPT can analyse students' language input and behaviour, and generate feedback that is tailored to their learning styles and needs. This can help teachers to provide more effective feedback that is targeted specifically to each student's strengths and weaknesses, and can therefore improve the overall quality of teaching and learning" (Source 10). Furthermore, ChatGPT has the potential to enhance or change the evaluation process for teachers (Source 4, 9, 10, 22, 26).

"Traditional education assessment primarily relies on exam scores and teacher evaluations, which often overlook students' actual learning situations and personalised needs. Future education assessments will be more diversified, utilising artificial intelligence technologies such as ChatGPT, to provide students with more objective and comprehensive evaluations. Students will be able to engage in targeted learning based on the assessment results" (Source 4). "For individual learners, once they possess a 'composite brain' consisting of the inner brain and 'outer brain' of ChatGPT, how do the two brains collaborate and work together? Does the strength of the outer brain make the inner brain stronger, or does it slowly atrophy? Therefore, teachers must rethink current evaluations and develop new thinking for learning evaluations that are in line with the synergy and symbiosis of the two brains" (Source, 9).

Despite the creative integration of ChatGPT into teaching, ethical issues of using the technology in the teaching process were most emphasised in the selected articles. Although ChatGPT can bring new tools and technologies to teachers in the field of education, it is essential to carefully weigh its pros and cons and take appropriate measures to mitigate any adverse effects. Due to the notable deficiencies of ChatGPT, many articles indicated that the rise of general large-scale models and their rapid penetration into education may also exacerbate ethical and security risks (Source, 3, 15, 16). It was suggested that the integration and development of artificial intelligence and education must adhere to a people-oriented approach, focus on cultivating students' high-level thinking, and establish a sound system of corresponding ethical norms (Source, 16). Some ethical and moral issues need to be taken seriously and addressed in the application of ChatGPT in the classroom. It was deemed necessary to regulate and constrain the technical design and application, as well as enhance education and awareness of user safety and data privacy (Source, 15).

RQ 3. How may ChatGPT change teachers' relationships with students?

The use of ChatGPT may change the way teachers interact with students (Source 4, 5, 6, 7, 9, 10, 11, 19, 20, 21, 23, 26). Among these articles examined, some articles mentioned that as teachers withdraw from the dominant role of learning and teaching, it would reduce the hierarchical relationship between teachers and students. Some pointed out that when knowledge becomes more easily accessible, soft skills and qualities development become increasingly important, and skills such as collaboration and creativity will be increasingly stressed. Accordingly, teachers will design more collaborative activities with students.

However, it was pointed out that the use of ChatGPT may also give rise to a trust crisis between teachers and students. "Once ChatGPT is introduced into schools and
students begin to use it for their own purposes, a composite brain is created. The existence of a composite brain can pose a certain academic integrity crisis for teachers when evaluating students' learning status" (Source, 9). In one of articles, it denoted that "cheating using ChatGPT not only undermines academic integrity but also makes it difficult for teachers to assess students' true proficiency, which is of no benefit to the students themselves. In addition to increasing the possibility of plagiarism, the content generated by ChatGPT may also contain biases, one-sidedness, and inaccuracies. This not only affects learning outcomes but also perpetuates stereotypes and negative views, exacerbating the spread of misinformation. Overreliance on artificial intelligence may also cause students to miss opportunities to develop critical thinking and enhance offline interpersonal communication skills. Therefore, it is not entirely unreasonable for some academic institutions to impose a ban on ChatGPT" (Source, 20).

5. Discussion

The findings of this study provide valuable insights into the potential impact of ChatGPT on the roles and identities of teachers. The use of ChatGPT in education creates opportunities for teachers to change their traditional teaching paradigm, explore innovative teaching pedagogies, and foster more collaborative and interactive relationships with students. By automating certain aspects of teaching, ChatGPT can help teachers to focus on higher-order thinking skills, such as critical thinking, problem-solving, and creativity, and to personalise learning experiences for individual students (Elgarf & Peters, 2022). On the other hand, the use of ChatGPT may also require teachers to shift their teaching responsibilities, leading to feelings of inadequacy or a loss of control over the teaching process. Teachers may feel that they are no longer the primary source of knowledge or expertise in the classroom, or that they are being replaced by a machine (Edwards & Cheok, 2018; Selwyn, 2019). Additionally, the integration of ChatGPT in education raises important ethical issues that have not yet been satisfactorily addressed. These issues include concerns over data privacy and security (Bertino et al., 2021), the potential for algorithmic bias and discrimination (Ryan, 2020), and the risk of plagiarism and academic dishonesty (Cotton et al., 2023). If these ethical concerns are not addressed, they may undermine the integrity and effectiveness of educational practices that rely on ChatGPT, and breed a trust crisis between students and teachers. While ChatGPT has the potential to transform and improve teaching and learning, it is important to address the potential risks associated with its use and to provide teachers with the necessary support and training to adapt to the changing landscape of education.

Given the complex and multifaceted impact of ChatGPT on the roles and identities of teachers, further research is needed to better understand the impact of ChatGPT on teachers. One direction for future research is to investigate the extent to which ChatGPT is contributing to teachers' identity crisis, and how this crisis can be mitigated through training and support. Another direction is to explore ways to rebuild trust between teachers and students in the context of AI-assisted teaching, such as providing greater transparency into the decision-making process of ChatGPT and prioritising human interaction and feedback in AI-assisted teaching. Finally, more research is needed to develop ethical guidelines for the use of AI in the process of teaching, including issues related to data privacy, algorithmic bias, and the potential for AI to perpetuate or exacerbate existing inequities in education.

This study would also inform future educational practices. As it is inevitable for human beings to embrace new technology, the study would inform the development of effective strategies for teachers to integrate this technology into their teaching. At the macro level, governments and academic associations should urgently develop relevant policies, regulations, and ethical norms for educational applications of generative artificial intelligence technologies like ChatGPT, and formulate specific, clear, and feasible policy recommendations (Schiff, 2022). At the meso level, schools or universities need to be aware of the identity crisis among teachers with ChatGPT taking part in roles of teachers and the trust crisis between teachers and students due to the deficiencies of Chat GPT in its transparency and reliability (von Eschenbach, 2021). Specifically, schools or universities should introduce AI detection tools for teachers to identify and differentiate between student
works and related materials generated by AI and those created by the students themselves to prevent a trust crisis between teachers and students. Meanwhile, they should provide relevant professional development programs on the ethical use of ChatGPT to prepare teachers technologically competent for future learning and teaching. At the micro level, teachers need to adjust their mindset, staying open yet critical towards the new technologies while integrating them into their learning and teaching.

Despite the valuable insights from this study, the findings should be viewed in light of its limitations. While ChatGPT emerges as a milestone in technological development, some countries, such as China, have banned the use of ChatGPT in education. Thus the data collected from articles and posts published on a single social platform may not entirely capture the impact of ChatGPT on teachers in other contexts. Moreover, the findings based on data from one social media platform may not generalise to the broader population of teachers because different platforms may have different features and norms that can affect the content and nature of the discourse representative. The WeChat platform is likely to attract specific types of users with particular interests, perspectives, and behaviours. Therefore, future studies can include data from a wider range of social media platforms for more comprehensive analysis. It is also advisable to include primary data from various contexts to examine how ChatGPT would impact teachers holistically.

References


Appendix 1

The identified sources that encompass three dimensions of teacher identity

<table>
<thead>
<tr>
<th>No.</th>
<th>Posts/Articles</th>
<th>Perceived self as a teacher</th>
<th>Pedagogical identity</th>
<th>Relationship with students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Source 1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Source 2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Source 3</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Source 4</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>Source 5</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Source 6</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>Source 7</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>Source 8</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Source 9</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>Source 10</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>Source 11</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Source 12</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Source 13</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Source 14</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Source 15</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Source 16</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Source 17</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Source 18</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Source 19</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>Source 20</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Source 21</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Source 22</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Source 23</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>24</td>
<td>Source 24</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Source 25</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Source 26</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Appendix 2

Coding scheme for examining the impact of ChatGPT on teacher identity

<table>
<thead>
<tr>
<th>Dimensions of identity</th>
<th>Subthemes</th>
<th>Contents or quotes from sources (examples)</th>
</tr>
</thead>
</table>
| Perceived self or role as a teacher | 1) Attitudes towards teaching occupation  
2) Traditional knowledge transmitter  
3) Taking on new roles  
4) Human teacher capabilities | e.g., While teachers can not be replaced by AI for the time being due to the difficulty for AI technology to replace human decision-making in complex situations (Source 4,6, 23) (Attitudes towards teaching occupation) |
| Pedagogical identity   | 1) Personalised learning  
2) Interactive learning  
3) Feedback  
4) Evaluation and assessment  
5) Ethics in technological pedagogy | e.g., As the functions of ChatGPT continues to evolve, teachers can utilise ChatGPT to generate a wide range of learning resources, such as online videos, games, virtual reality, and other forms, thereby increasing learning interest and motivation. (Source 4,7) (Interactive learning) |
| Relationships with students | 1) Flat or democratic relationships  
2) Trust crisis  
3) Reduced communication | e.g., "Once ChatGPT is introduced into schools and students begin to use it for their own purposes, a composite brain is created. The existence of a composite brain can pose a certain academic integrity crisis for teachers when evaluating students' learning status." (Source, 9) (Trust crisis) |
Designing Faculty Development Programs by a Team from Different Majors

Yukari KATOa*, Yukihiko YAMASHITAb, Hisashi HATAKEYAMAA & Toshiya OISHIB

a Center for Innovative Teaching and Learning, Tokyo Institute of Technology, Japan
b Science Academy, Chiba University of Commerce, Japan
*kathy@citl.titech.ac.jp

Abstract: This study investigated how three faculty members and an external lecturer selected training content and implemented teaching methods for faculty development (FD) to ensure learning opportunities for faculty members at science universities. Qualitative analysis revealed the following three points: (1) the lack of a clear image of the target participants causes discrepancies between participant needs and the FD seminar planners; (2) the faculty developer (FDer) is apt to “chase after two hares,” when seeking student-centered teaching style and instructors’ language skills improvement; and (3) FD seminars should be focused on a specific topic to be more attractive to the target participants.

Keywords: Professional Development, English Medium Instruction, Faculty Developer, Steps for Coding and Theorization

1. Introduction

Internationalization has been vigorously promoted in Japanese higher education through government-led policies such as the Global 30 Project and the Top Global University Project (MEXT, 2014). Japan’s Ministry of Education, Culture, Sports, Science and Technology (MEXT) has been providing support to improve the international competitiveness of Japanese higher education through its “Super Global University (SGU)” program, which focuses on 13 top universities (Type A) that conduct world-class education and research, and 24 globalization-driving universities (Type B) have been leading the globalization of Japanese society since 2014. The selected universities have provided their students with globalization-oriented education and research by fostering students’ global responsiveness and further accelerating exchanges and collaborations with world-class universities. SGU-adopted schools; however, they must meet the goals established in “common outcome indicators,” such as international students as a percentage of all students, the rate of courses taught in foreign languages, a flexible academic calendar (introduction of the quarter system), and various organizational and curricular reforms. To achieve these indicators and demonstrate their internationalization initiatives worldwide, top universities held workshops in English by invited lecturers from leading overseas universities and launched joint research projects in cooperation with overseas partner universities. Furthermore, English-medium instruction (EMI) is encouraged to attract talented international students and researchers, and both existing and newly hired faculty members are increasingly required to “teach in English.” At the Tokyo Institute of Technology, more than 90% of graduate school courses are taught in English to comply with the SGU outcome indicators. The Center for Innovation in
Teaching Language (CITL), to which the authors belong, has invited lecturers from overseas partner universities (University of Queensland) and English-language institutes in Japan to hold training sessions to promote the use of English in class.

2. FD Seminar with Flipped Learning Material

Preparation for FD Seminar

The theme of the FD training in FY2022 was “Acquisition of knowledge and skills for practical graduate school lectures in English (EMI).” As shown in Table 1, preparation for the FD training started in April 2022. Faculty member A was in charge of the FD training content, creating teaching materials, and conducting the training with an external Instructor B.

Table 1. FD Training Project Members and Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Specialty</th>
<th>Age, Years at CITL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty A</td>
<td>Planning FD seminar</td>
<td>Educational Technology</td>
</tr>
<tr>
<td>Lecturer B</td>
<td>Lecture</td>
<td>English Teaching</td>
</tr>
<tr>
<td>Faculty C</td>
<td>Supporting FD seminar</td>
<td>Informatics</td>
</tr>
<tr>
<td>Faculty D</td>
<td>Supporting FD seminar</td>
<td>Science Education</td>
</tr>
</tbody>
</table>

2.1 Team for FD Seminar

Initially, only Faculty Member A and external Lecturer B, who were in charge of the training, consulted with each other and decided on the FD training content. For example, the range of sessions on (a) the concept of lesson planning, (b) assignments and homework, and (c) how to start a class were planned as English lecture basics. Additionally, two lectures on strategies to promote active learning among students were given at different language levels to accommodate various learning styles as interactive sessions. They will also touch on a class environment in which students can easily ask questions to stimulate interaction among teachers and students and methods for verifying results (Kato & Oishi, 2022 a; 2022b). To design an effective FD training program to promote specialized education in English, internal meetings were held not only including Faculty A and external Lecturer B, who are foreign language education specialists but also other full-time CITL faculty members (C and D) who have experience in specialized higher education in science. They discussed “what kind of training would be acceptable to science faculty” and “what kind of content would interest science faculty who are busy with research, enabling them to participate without creating a burden” with Faculty C, who has experience teaching specialized fields in English, and Faculty D, who had attended classes in his university’s doctoral program. Based on the results of this discussion, the person in charge of planning (Faculty A) developed the FD courses and publicized them within the university. Despite multiple publicity campaigns, however, fewer than 10 participants attended. Unfortunately, the project was not of interest to science faculty members.

2.2 Research Question

This study aims to clarify the decision-making process of planning and operating FD training programs for EMI courses and develop knowledge for effective planning and implementation of FD training at science universities with a strong research orientation. In particular, the authors sought to elucidate how a newly appointed Faculty Developer
(FDer, Faculty A) searched for clues to establish FD training tailored to universities with solid research activities through discussions with other members (Faculty C and D). The question we have to ask here is what kind of FD training content and implementation methods are acceptable for faculty at science universities. In addition, this issue examined building collaborative relationships among faculty members and problem-solving teams, which were discussed in previous studies (Kato et al., 2023a; 2023b).

3. Research Method

3.1 Participants

The subjects of this study were three faculty members at a science university and one external lecturer. As shown in Table 1, Faculty members A and D were newly transferred from other universities. Hence, they needed more information on the roles of faculty members in research-oriented science education at their home universities and how to manage graduate classes using English. Faculty A was in charge of this training course, and Faculty C and D belonged to the same organization. Faculty A had approximately ten years of experience as an FDer, and Faculty D had previous experience as a professor at the information processing center of another university and has been involved in educational improvement and evaluation at the Center for educational technology since last year. Alternatively, Faculty C is an experienced researcher who transferred from a specialized department to an education-related center within the university. He has teaching experience at the undergraduate and graduate levels in his area of expertise and is familiar with the university’s situation.

3.2 Qualitative Analysis Method (Steps for Coding and Theorization)

In this study, the authors attempt to explore and structure the issue of FD training in science universities using Steps for Coding and Theorization (SCAT), a qualitative research method, to explore how the “desirable training style” was determined in small-group discussions. The authors then conducted a qualitative analysis using SCAT proposed by Otani (2011) to clarify the process of discussion among the trainers on how to proceed with “FD training for promoting English language teaching at science universities.” SCAT is an analytical method providing a roadmap for discovering potential meanings in language records collected through observation and interviews and for developing new concepts. In addition, qualitative research methods aimed at hypothesis generation are considered suitable for exploratory investigations to obtain helpful information to overcome the current situation (Fukushi & Nago, 2011; Kato et al., 2023a; 2023b).

3.3 Data Analysis

The first author transcribed the audio recordings. The transcripts were analyzed using SCAT, a sequential, thematic qualitative data-analysis technique (Otani, 2008; 2011). It includes coding steps from open to selective, storyline creation using final selective codes, and creating theories from the storyline. Accordingly, each utterance of faculty members was considered to be a single recording unit. These recording units were then classified into categories and subcategories based on the similarity of their semantic content, with different themes extracted following the exercise. This approach was selected for its explicit analysis process, which integrates qualitative data analysis with theoretical coding, and for its efficiency and validity of theorization using relatively small-scale data (Otani, 2008). This analysis covered the language records of a December 19, 2022, meeting. The recording time was 53 minutes and 12 seconds. Faculty A was the meeting facilitator, and three faculty members expressed their opinions on “In what type of FD training course content and implementation would science faculty members be
interested in participating?” according to the interview guide in Table 2. The verbal recordings of the meetings were divided into segments based on the SCAT analysis procedure, resulting in 301 segments. Some segments (8 segments, 28 seconds) were excluded from the analysis because they included reports of FD training at other institutions attended by Faculty D. There were 293 total segments.

Table 2. Interview Guide

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>How to teach a specialized class in English at a science university.</td>
</tr>
<tr>
<td>2</td>
<td>How to promote active learning and other student-centered classes.</td>
</tr>
<tr>
<td>3</td>
<td>What kind of FD training content and implementation methods are acceptable for faculty at science universities?</td>
</tr>
<tr>
<td>4</td>
<td>What is the role of an FDer? Did she promote participants' reflections?</td>
</tr>
</tbody>
</table>

4. Results

Qualitative analysis identified the main categories and subcategories in Tables 3 and 4. Examples of each aspect of the nine main themes are provided below, with the subcategories and main categories denoted by angle brackets (<…>) and square brackets ([…]), respectively. In the following paragraphs, double quotation marks (“…””) denote representative descriptions, and the numbers inside brackets denote unit numbers for each participant in Table 1. At the preparatory meeting for the FD training on December 19, 2022, faculty members at CITL expressed their opinions on the “ideal FD training program” based on their own experiences and the situation of their organizations.

The research question “In what type of FD training course content and implementation would science faculty members be interested in participating?” prompted us to aggregate five itemized subcategories into two main categories: [FDer Losing the Way] and [Clear and Simple Course]. As shown in Table 3, the first main category included the following three subcategories:<Targets are blurred>, <Pursuing two hares>, and <Discrepancies in training content>. The second main category [Clear and Simple Course] clarified that FD training should be <clear vision of training courses> or deal with <Simple Skill Training> that focuses on language skills, as shown in Table 4.

Table 3. Subcategories of Problems with FD Training Courses

<table>
<thead>
<tr>
<th>[FDer Losing the Way]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Targets are blurred&gt;</td>
<td>I had the idea that I should be teaching more about teaching strategies, but Instructor B was more interested in discussions, how to teach (language activity), icebreaking, holding a clinic, and so on. B was more focused on activities, such as how to teach discussion, icebreaking, etc. (186: A).</td>
</tr>
<tr>
<td></td>
<td>If you put two things in, you end up with a different target audience, and It’s a bit of a blur, and I think it is more for people who are already using English or are thinking about using it, or know how to use it. (234:D).</td>
</tr>
<tr>
<td>(omission)</td>
<td>It might be an overstatement to say that this is a bit of a niche, but we are aiming at that audience, so it’s not a bad direction to take. It’s not a bad direction to take, and it’s excellent. It’s not a bad direction to take. (270:D).</td>
</tr>
</tbody>
</table>

| <Pursuing two hares>  | (They may not go to FD training.) They may only be able to make it to this level if they learn how to communicate well, speak, explain, and give presentations in English, which may be an overstatement. You Can’t reach this level unless the participants clear the primary level. (218:D) |

Table 4. Clear and Simple Course

<table>
<thead>
<tr>
<th>SFer Mongling Way</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SFer Mongling Way</td>
<td></td>
</tr>
<tr>
<td>SFer Mongling Way</td>
<td></td>
</tr>
<tr>
<td>SFer Mongling Way</td>
<td></td>
</tr>
</tbody>
</table>
I wonder how it would work in this situation. So, we should have some specific specialized lectures. I do not know. I do not care if it is a video or something, but he (Lecturer B) does not specialize in that (science) field. (188:C)

I don’t think it would be a good fit (to bring lesson plans) to have them do a mock class and hold a clinic. It might be possible to do what I mentioned earlier. It will still be okay if it is the kind of group-work lecture I mentioned earlier. (192:C)

Table 4. Subcategories of Ideal FD Training Courses for a Science University

<table>
<thead>
<tr>
<th>Subcategories of Ideal FD Training Courses</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Clear Vision of Training Courses&gt;</td>
<td>Yes, it would be nice to have a demonstration. So, it would be nice to have a demonstration. Like in the sciences, we are doing that kind of interaction with math or something like that. (23:C).</td>
</tr>
<tr>
<td></td>
<td>One example of a non-native but foreign-style class based on current trends is a class in which students participate. If we connect this to a discussion of how this type of class style could be introduced, it would clarify what the teacher learns. Since it is pretty clear what the teachers will gain from this FD training, even considering the time constraints, it would be better to focus on such a specific topic and not make it too much. (121:D)</td>
</tr>
<tr>
<td></td>
<td>That is not the point; that is not possible. I am sorry. Let us give up on that, but we do have group work, so I think we should try to improve the English group works. I am sorry, but I cannot do that. (229:C)</td>
</tr>
<tr>
<td>&lt;Simple and Focusing Training&gt;</td>
<td>I cannot say. Many things can be said in Japanese but not in English. (57:C)</td>
</tr>
<tr>
<td></td>
<td>I have been working on this since April or May, and recently I have been reflecting on whether I overdid it a bit. I could have done it sooner if I had used a better lecture method like that of English-language institutes. (209: A).</td>
</tr>
</tbody>
</table>

That should not be possible. I am sorry. I have given up on that, but we have group work to improve the English group work. I am sorry, but I am afraid that is not possible. (229:C)

1. Conclusion
This qualitative and exploratory study intended to define how faculty developers at a CITL design practical training courses for professors and how they would adjust the plan to meet participants’ needs and the situation at science universities.

First, in preparation for this FD training course, the FDer led the FD training to promote EMI in specialized classes at science universities and asked an external lecturer, Lecturer B, to prepare teaching materials and conduct the training. However, Faculty A and Lecturer B, who specialize in foreign language education, did not fully understand the specialized education at science universities. Therefore, a further discussion involving Faculty C and D regarding suitable FD training was conducted, which organized a team from different major fields.

Second, Faculty A reflected on her own struggle to design and implement successful FD training courses at a science university. Although she was eager to implement good FD courses at a science university, she did not find the solutions
convincing. For example, Faculty A stated, “Instructor B was more interested in discussions, how to teach (language activity), icebreaking, holding a clinic, and so on” (unit 186). In the meeting, Faculty C was hesitant to set the hard-burden courses under the strong recommendation and pressure to eager in high-level research, saying, “I don't think it would be a good fit (to bring lesson plans) to have them do a mock class and hold a clinic” (unit 192).

Finally, the research question prompted us to aggregate five items into two main categories: [FDer Losing the Way] and [Clear and Simple Course]. The first main category included the following three subcategories: <Targets are blurred>, <Pursuing two hares>, and <Discrepancies in training content>. As shown in Table 3, for newcomer Faculty A (FDer) at CITL, the target audience and content for FD training were unclear at a science university. She then planned the “dual pursuit” FD seminar to improve both educational methodology and English language skills. However, this has created a mismatch with the "training needs" of science faculty. These three subcategories are summarized as "FDer Losing the Way." It was suggested that training be reconstructed as a [Clear and Simple Course]. It was also clarified that FD training should have a <clear vision of training courses> or deal with <Simple Skill Training> that focuses on only language and classroom communication skills. These two subcategories were conceptualized as tips for FD training course improvement.

There were limitations to conducting an analysis of only one discussion, the findings of which might be helpful for future qualitative studies that would require clarity and requirements for successful FD courses at science universities.

Acknowledgments
JSPS KAKENHI Grant Number JP23K02682 has funded this project. Information about this project can be found online at http://www.katoyukati.net/

References

Ministry of Education, Culture, Sports, Science, and Technology, "Status of Reform of Educational Contents, etc. at Universities in the 2023 Academic Year (Summary)," https://www.mext.go.jp/content/20211104-mxt_daiagakuc03-000018152_1.pdf (referenced on 2023.2.25)


Otani, H. (2008). Proposal of SCAT, a qualitative data analysis method with 4 step coding: A theorizing procedure easy to start and applicable to small data", Graduate School of Educational Developmental Sciences, Nagoya University Bulletin (Educational Science), 54(2), 27–44.

Abstract: This paper presents study results regarding digital tools and work methods which early childhood pre-service teachers consider using in their future pedagogical activities with children. These results are part of wider international research regarding the digital competencies of the students-future early childhood teachers. The purpose of the presented results is to compare students’ attitudes towards digital tools for working with kids from the perspective of 8 countries (Norway, The UK, Portugal, Slovenia, Poland, Ukraine, Turkey, and Jordan). There were 772 respondents – students of the Early childhood pedagogical specialization. The study follows the design of a survey based on the Theory of Action. The results of the study revealed that the pre-service early childhood teachers of all nations participating in the study presented positive approach towards the use of IT tools in kindergarten. The digital tools which were indicated by students as those that will be most often used in future work are the Internet as a source of knowledge, digital programs for communication with parents, video, pedagogical apps, and smart boards. Differences in answers between representatives of different nations can be explained by the specificity of the functioning of educational systems in different countries as well as differences in Early Childhood teachers’ curriculums.

Keywords: Digital tools, kindergarten, international study, early childhood pre-service teachers

1. Introduction

Professional digital competence for early childhood educators (ECE) is important because digital devices for young children are evolving at a rapid rate (Arnott et al., 2018). Nevertheless, researchers and society, in general, seem to be more conflicted and critical towards the use of technology in ECE compared with other educational levels. Some researchers highlight that a ‘moral panic’ surrounding screen time appears to have positioned digital technology as an enemy of early childhood practice - one that could hinder
the implementation and development of digital practices in early education (Fleer, 2020). On the whole, early childhood teachers’ perceptions of the benefits and disadvantages of educational technology integration have been the subject of a moderate amount of empirical research (Hatzigianni & Kalaitzidis, 2018; Magen-Nagar I & Firstater, 2019). Therefore, the current study investigates pre-service early childhood teachers’ attitudes towards the use of technology in the educational setting. The goal is to gain a deeper understanding of this issue within an international educational context.

2. Materials and Methods

2.1 Study sample

This paper presents an overview of the perspectives of early childhood pre-service teachers (n=772) regarding the use of different digital tools and educational methods within their professional practices. The survey was conducted at educational institutions in eight different countries during the academic year 2021-2022. The universities which were included in the study: UiT the Arctic University and NLA University from Norway College (n=191); University of Primorska, Slovenia (n=177); ISEC Lisboa, Portugal (n=93); Pedagogical University of Krakow, Poland (n=83); Erciyes University, Turkey (n=82); H.S. Skovoroda Kharkiv National Pedagogical University, Ukraine (n=55); Birmingham City University, UK (n=46) and Al-Hussein Bin Talal University, Jordan (n=45). Data were collected from institutions providing teacher education in each country, applying the same survey structure for all the institutions. The overall average response rate was 78.38 %.

2.2 Research questions

1. What are the pre-service early childhood teachers' attitudes towards the use of IT tools in kindergarten differed in every country?
2. What kind of digital tools are ECE students in different countries more likely to be expected to use in their future practices?

2.3 Study Tools and Procedure

The results presented in this paper are a part of wider international research regarding digital competencies of the students-future early childhood teachers.

The survey includes pre-service ECE teachers' attitudes towards the future use of digital technology in kindergarten. The following items of the research tool were extracted for this variable: ECE students’ self-reflection on the use of digital tools in future pedagogical work in kindergartens; availability of digital tools in early childhood education; difficulties with providing digital tools in kindergarten for economic reasons; the use of digital technology because of professional responsibility; frustration at the need to use IT tools in kindergarten. This part of the survey consists of optional questions comprising a range of 5-point Likert-scale statements ranging from 1 (Strongly disagree) to 5 (Strongly agree). The item regarding the digital tools which students will use for educational activities with children contains a list of various digital tools and closed answers to indicate the frequency of their use (1 - never, 2 - rarely, 3 - occasionally, 4 - often and 5 - extensively).

An English version of the survey was used as a template for translating the survey into the different languages needed. Translations were done by researchers from each nation, thus ensuring a high level of accuracy. Data were collected through the use of online survey tools, mainly through Nettskjema, a Norwegian tool for the design and implementation of surveys and secure online data collection, made for the university and college sector.

Considering the ordinal character of the measurement of variables, the between-group comparisons were conducted using the Mann-Whitney U and Wilcoxon W nonparametric tests.
4. Results

4.1 Pre-service ECE teachers’ attitudes towards future use of digital technology in kindergarten

The results of the single items regarding the pre-service ECE teachers’ attitudes towards future use of digital technology in kindergarten revealed differences among the different nations. These differences are presented in Table 1, using mean scores and standard deviations.

Table 1. Mean Scores for Single items regarding the pre-service ECE teachers’ attitudes towards future use of digital technology in kindergarten

<table>
<thead>
<tr>
<th>Country</th>
<th>Self-reflection on the use of digital tools in future pedagogical work in kindergartens</th>
<th>Availability of digital tools in early childhood education</th>
<th>Difficulties with providing digital tools in kindergarten for economic reasons</th>
<th>Use of digital technology as professional responsibility</th>
<th>Frustration at the need to use IT tools in kindergarten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>4.12 (0.99)</td>
<td>3.96 (1.02)</td>
<td>4.16 (0.91)</td>
<td>2.16 (1.02)</td>
<td><strong>2.30 (1.11)</strong></td>
</tr>
<tr>
<td>UK</td>
<td>4.26 (0.58)</td>
<td>3.76 (0.85)</td>
<td>4.04 (0.76)</td>
<td>2.48 (0.84)</td>
<td>2.63 (0.74)</td>
</tr>
<tr>
<td>Portugal</td>
<td>4.27 (0.51)</td>
<td><strong>4.12 (0.69)</strong></td>
<td><strong>4.31 (0.59)</strong></td>
<td>2.21 (1.30)</td>
<td>2.45 (1.23)</td>
</tr>
<tr>
<td>Slovenia</td>
<td><strong>3.56 (0.82)</strong></td>
<td>3.33 (1.05)</td>
<td><strong>3.69 (0.83)</strong></td>
<td>2.42 (0.95)</td>
<td>2.84 (0.96)</td>
</tr>
<tr>
<td>Poland</td>
<td>4.13 (1.11)</td>
<td>3.65 (1.08)</td>
<td>3.77 (1.03)</td>
<td>2.83 (1.24)</td>
<td>2.34 (1.11)</td>
</tr>
<tr>
<td>Ukraine</td>
<td>4.00 (0.64)</td>
<td>3.67 (1.07)</td>
<td>4.04 (0.86)</td>
<td>2.42 (0.98)</td>
<td>2.75 (1.02)</td>
</tr>
<tr>
<td>Turkey</td>
<td><strong>4.32 (0.58)</strong></td>
<td>3.95 (0.87)</td>
<td>4.05 (0.72)</td>
<td><strong>2.00 (0.96)</strong></td>
<td>2.52 (0.92)</td>
</tr>
<tr>
<td>Jordan</td>
<td>3.69 (0.85)</td>
<td>3.53 (1.14)</td>
<td>4.07 (0.99)</td>
<td><strong>3.58 (0.72)</strong></td>
<td><strong>3.20 (1.10)</strong></td>
</tr>
<tr>
<td>average</td>
<td>4.04 (0.76)</td>
<td>3.74 (0.97)</td>
<td>4.02 (0.84)</td>
<td>2.51 (1.00)</td>
<td>2.63 (1.02)</td>
</tr>
</tbody>
</table>

Mann-Whitney U: 4204.50
Wilcoxon W: 19957.50 20433.00 20545.50 3833.00 20821.50
Z statistic: -6.09 -6.12 -6.11 -7.45 -4.54
Asym. sig. (2-tailed): < 0.001 < 0.001 < 0.001 < 0.001 < 0.001

Note: \( r_g \) - Effect size between the mean ranks in the two compared groups;
The bolded values concern the groups in which the lowest and highest scores relate to a specific attitude component.

When looking at nations’ mean scores independently, the results tend to vary. The effect sizes between the lowest and highest scoring nations are significant for all the items. The biggest difference between the highest and lowest score relates to the use of digital technology due to professional responsibility \((r_g=0.77)\). Teachers from most countries disagree...
and are neutral regarding the statement of the need to use IT as part of professional duties. Teachers from Jordan, however, are generally in favor of such a justification for the use of IT. For the other components of the attitude the effect sizes were similar (about $r_2=0.40$). The highest level of self-reflection on the use of digital tools in future pedagogical work in kindergartens were recorded among students from Turkey (4.32), whereas the lowest one were observed among students from Slovenia (3.56). The results for the other attitude components should be interpreted similarly (see Table 1).

4.2 Students’ future Early Childhood teachers’ opinions about the digital tools for use during ECE from the perspective of different countries

This variable is a good indicator of the respondents’ reported future use. Obtained results are based on the mean scores for 17 different tools and work methods (see Table 2).

<table>
<thead>
<tr>
<th>Mean Score (SD)</th>
<th>Norway (1,07)</th>
<th>UK (0.94)</th>
<th>Portugal (0.89)</th>
<th>Slovenia (0.78)</th>
<th>Poland (1.00)</th>
<th>Ukraine (1.43)</th>
<th>Turkey (0.80)</th>
<th>Jordan average (1,06)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quiz</strong></td>
<td>2.66 (0.90)</td>
<td>3.04 (1.23)</td>
<td>3.27 (0.80)</td>
<td>2.91 (0.80)</td>
<td>3.05 (1.00)</td>
<td>3.64 (1.23)</td>
<td>2.79 (1,05)</td>
<td>3.64 (1,02)</td>
</tr>
<tr>
<td><strong>Digital tools for presentation</strong></td>
<td>3.06 (1,02)</td>
<td>3.78 (1,06)</td>
<td>3.01 (0.89)</td>
<td>3.38 (1.00)</td>
<td>3.94 (1.06)</td>
<td>3.60 (0.89)</td>
<td>3.37 (1,05)</td>
<td>3.38 (1,02)</td>
</tr>
<tr>
<td><strong>Programs to create texts</strong></td>
<td>3.56 (1,21)</td>
<td>3.70 (0.94)</td>
<td>3.15 (0.99)</td>
<td>3.84 (0.96)</td>
<td>4.00 (1.06)</td>
<td>4.07 (0.81)</td>
<td>2.82 (1,00)</td>
<td>2.98 (1,19)</td>
</tr>
<tr>
<td><strong>Video use</strong></td>
<td>3.49 (1,07)</td>
<td>4.07 (0.55)</td>
<td>4.04 (0.82)</td>
<td>3.69 (1.12)</td>
<td>2.20 (0.84)</td>
<td>3.76 (1.04)</td>
<td>4.02 (1,10)</td>
<td>3.69 (1,04)</td>
</tr>
<tr>
<td><strong>Production of film/video/animation</strong></td>
<td>3.05 (1,10)</td>
<td>3.72 (0.94)</td>
<td>3.67 (0.86)</td>
<td>3.27 (0.99)</td>
<td>2.22 (1.21)</td>
<td>3.24 (0.80)</td>
<td>3.89 (1,02)</td>
<td>3.53 (1,12)</td>
</tr>
<tr>
<td><strong>The Internet as a source of knowledge</strong></td>
<td>4.17 (0,91)</td>
<td>4.24 (0,79)</td>
<td>3.95 (0,80)</td>
<td>3.81 (1,07)</td>
<td>4.69 (0,66)</td>
<td>4.31 (0,81)</td>
<td>4.16 (0,78)</td>
<td>3.98 (0,92)</td>
</tr>
<tr>
<td><strong>iPad</strong></td>
<td>3.78 (1,03)</td>
<td>3.70 (1,05)</td>
<td>2.92 (0,73)</td>
<td>2.10 (1,03)</td>
<td>2.07 (1,30)</td>
<td>2.35 (1,12)</td>
<td>3.20 (1,16)</td>
<td>2.58 (1,27)</td>
</tr>
<tr>
<td><strong>Pedagogical apps</strong></td>
<td>3.79 (0,96)</td>
<td>3.70 (0,77)</td>
<td>3.68 (0,77)</td>
<td>3.30 (0,99)</td>
<td>3.53 (1,17)</td>
<td>3.93 (0,86)</td>
<td>4.33 (0,77)</td>
<td>4.00 (0,93)</td>
</tr>
<tr>
<td><strong>Toys and technology for coding</strong></td>
<td>2.73 (1,17)</td>
<td>3.37 (0,88)</td>
<td>3.05 (0,95)</td>
<td>2.45 (0,99)</td>
<td>1.60 (1,01)</td>
<td>2.89 (1,05)</td>
<td>3.70 (1,05)</td>
<td>2.62 (1,17)</td>
</tr>
<tr>
<td><strong>Digital camera</strong></td>
<td>3.64 (1,07)</td>
<td>3.46 (1,04)</td>
<td>3.41 (0,78)</td>
<td>3.58 (1,00)</td>
<td>2.71 (1,43)</td>
<td>3.15 (0,93)</td>
<td>3.39 (1,02)</td>
<td>2.56 (1,06)</td>
</tr>
<tr>
<td><strong>Digital microphone</strong></td>
<td>2.49 (1,18)</td>
<td>2.89 (1,04)</td>
<td>2.77 (0,98)</td>
<td>2.44 (0,94)</td>
<td>1.39 (0,93)</td>
<td>2.27 (0,99)</td>
<td>3.24 (1,02)</td>
<td>2.51 (1,13)</td>
</tr>
<tr>
<td><strong>Digital storytelling</strong></td>
<td>3.33 (1,07)</td>
<td>3.48 (0,80)</td>
<td>2.84 (0,80)</td>
<td>3.10 (1,05)</td>
<td>1.99 (1,21)</td>
<td>3.00 (1,02)</td>
<td>3.49 (1,01)</td>
<td>3.07 (1,12)</td>
</tr>
<tr>
<td><strong>Digital music and sounds (as a producer)</strong></td>
<td>3.30 (1,3)</td>
<td>3.20 (1,00)</td>
<td>3.29 (1,02)</td>
<td>3.22 (1,11)</td>
<td>1.55 (0,91)</td>
<td>3.25 (1,14)</td>
<td>3.66 (0,84)</td>
<td>3.00 (1,23)</td>
</tr>
<tr>
<td><strong>Digital music and sounds (as a consumer)</strong></td>
<td>3.79 (1,14)</td>
<td>3.63 (0,77)</td>
<td>3.65 (0,80)</td>
<td>3.87 (0,98)</td>
<td>2.72 (1,49)</td>
<td>3.58 (0,99)</td>
<td>3.83 (0,81)</td>
<td>3.36 (1,28)</td>
</tr>
</tbody>
</table>

Table 2. Mean Scores and standard deviation (SD) for Single Variable regarding digital tools and work methods which will be used by pre-service early childhood teachers in their future pedagogical activities with children
The obtained results show that the Internet as a source of knowledge was indicated by early childhood education students from all the nations. Students from Norway reported that they intend to use digital programs for communication with parents most often, similarly indicated by students from Turkey and Jordan. Video, as a tool for frequent uses in kindergarten, was picked up by students from the UK, Portugal and Turkey. Students from Turkey and Jordan also indicated their intention to use Pedagogical apps frequently in their future professional work. Students from the UK were the only ones who indicated that they intended to use the Smartboard often in their work with children in kindergarten.

Summarizing, there were a few kinds of digital tools indicated by students as those that will be most often used in future work, including the Internet as a source of knowledge, digital programs for communication with parents, video, pedagogical apps and smart boards. Differences in answers between representatives of different nations can be explained by the specificity of the functioning of educational systems in different countries as well as differences in Early Childhood teachers’ curriculums.

5. Discussion

The obtained results indicate a tendency towards positive attitudes regarding the use of IT tools in kindergarten demonstrated by pre-service early childhood teachers of all the countries participating in the study. This result indicates the high credibility of the implementation of innovative technologies in the educational process of kindergarten students after starting their professional career in kindergarten. Our results are in line with some previous ones regarding mobile learning as a kind of education in kindergarten (Kalogiannakis and Papadakis, 2019). The authors stated that the pre-service teachers’ attitudes towards the usefulness of mobile learning in the teaching process has the strongest influence on their intention to adopt mobile learning followed equally by perceived ease of use. According to the recent studies (Zilka, 2021) the attitudes of preservice teachers toward integrating computers into kindergartens are mainly positive, but their perception of the situation as stressful and threatening is higher than among in-service teachers.

Our results regarding the kinds of digital tools which the pre-service early childhood teachers are going to use in their future teacher activities in kindergarten are quite diverse across nations. All the respondents pointed to the Internet as a source of knowledge. Digital programs for communication with parents, video, pedagogical apps and smart boards were also mentioned as students’ intentions to use during future teaching with kids. This result shows ECE students’ readiness for the integration of some digital tools into the educational process in kindergarten. Similar results were reported by Chinese authors, who highlighted that, preservice teachers conceptualize ICT as screen-based technologies such as interactive whiteboards, desktop and laptop computers, e-projectors and smartphones, as well as educational software and applications that can be used on these devices (Dong & Mertala, 2021).

6. Conclusion

Overall, the dynamics of digital practices across the countries reveal that pre-service early childhood teachers of all the nations participating in the study presented positive
intentions towards the use of IT tools in kindergarten. Respondents from all the countries are aware of different digital tools, hardware, and software for use in ECE. The digital tools which were indicated by the students as those that will be most often used in future work are the Internet as a source of knowledge, digital programs for communication with parents, video, pedagogical apps and smart boards. This outcome calls for a closer examination to determine whether certain factors can explain the different nations’ strong explanatory power and statistical significance in terms of attitudes and skills as predictors. Related to this, qualitative studies of different educational contexts are needed to understand the complexity of our results. Finally, further research into the emerging digital pedagogies in ECE is needed so that there can be more discussions and development of education built on knowledge and research.

The scope of this study was to present how pre-service early childhood teachers of different countries consider using digital tools and work methods in their future pedagogical activities with children. The findings of the previous studies we examined, showed that kindergarteners can benefit from the use of new technology and can gain many advantages and profits using the appropriate educational tools that offer the use of ICT.

However, the teachers’ intentions for using different digital tools are thought to be significant and demand elaboration and clarification, especially from the perspective of comparative international studies. There are lot of learnings that come with the execution of an multi-institutional and multi-national study like addressing cultural differences, ethical review constraints, technology policy differences etc. Address these challenges will offer valuable insights to researchers in the Asia-Pacific region.

7. Limitations

First of all, there was a small number of respondents and the study was conducted only once and it was cross-sectional. Thus, more research is needed, carried out in more advanced designs. Finally, a significant challenge was the investigators’ cultural differences and the need to adapt the research tools to the participants’ national languages with maintaining satisfactory psychometric characteristics.

References


Democratising AI education: Teaching autoencoders to out-of-school children from low-income backgrounds

Saumay GARG

Department of Computer Science and Information Technology, Jaypee Institute of Information Technology, India
*saumaygarg@gmail.com

Abstract: Artificial Intelligence (AI) has become an integral part of personal computing devices and is gaining importance in school curriculums. Several studies and tools have been developed for teaching AI to school-aged children and it continues to be a growing area of research in computer science education. However, most of these interventions are targeted towards the more privileged and therefore cannot be directly used with children from socio-economically backward families. Children from such backgrounds are often unable to complete even basic elementary education due to several social, economic and systemic challenges. This work aims to help extend the reach of AI education to these less privileged children. A learning activity based on the five AI4K12 big ideas is presented, the activity follows a series of sessions aided by an easy to create tool for supporting tinkerability. The activity attempts to give a high-level overview of the functioning of autoencoders. The feasibility of this approach is verified with a case study involving out-of-school children from low-income families. The findings show that all of the AI4K12 five big ideas can be addressed and effectively taught to non-English speaking children with low numeric literacy. The proposed approach was also shown to increase children’s curiosity and sense of agency while learning about AI.

Keywords: Computational Thinking, Inclusive Education, Artificial Intelligence

1. Introduction

Artificial Intelligence (AI) has been accepted to be an integral part of computational thinking (Denning and Tedre, 2019). This is primarily because AI has become omnipresent with most computing devices relying on its applications in one way or another. However, AI can also have inherent biases that lead to discrimination towards marginalized groups (Angwin et al, 2016; Buolamwini et al, 2018). This omnipresence and potential for bias requires that children are taught about AI in schools in both a technical and sociological context.

Artificial Intelligence for K-12 (AI4K12) is a globally recognized initiative that supplies practitioners with standardized guidelines for AI education in schools. This initiative was started as a joint project by the Computer Science Teachers Association (CSTA) and the Association for the Advancement of Artificial Intelligence (AAAI). The AI4K12 guidelines are developed around the five big ideas for AI education. These big ideas broadly describe the goals for K-12 AI education while also serving as guiding principles for developing related resources and best practices (Touretzky et al, 2022). The 5 big ideas help ensure that the curriculum and resources focus on teaching all significant areas encompassing the AI field and not just a few superficial concepts. These five ideas are represented in Figure 1.

Smart phones have become extremely common in recent years, so much so that even marginalized and low-income families own smartphones now. According to a recent report, up to 88% of Indian households own a smartphone (Bhattacharya, 2022). And surprisingly, even children from these socio-economically backward backgrounds are extremely proficient in the usage of these tiny mobile computers. But these kids are not educated in the more technical functioning of computers and AI. Such children are very likely to be out of school, have never
attended school or have subpar resources in their poorly funded schools. To put the magnitude of this problem in context, around 59.1 million primary school-aged children don’t attend schools (UNESCO, 2019), primarily due to financial and social constraints. Consequently, these families continue to be stuck in an endless cycle of poverty and are unfortunately devoid of the opportunities that education and computer skills can bring. Lack of research into teaching practices for such populations makes the educational disparity even more concerning.

Research into teaching and learning of AI for K-12 has examined various best practices and studied the efficacy of teaching resources for AI. Some of these studies include works by Druga et al (2019), Long et al (2020) and Payne et al (2019). Though the progress in research for supporting AI literacy is promising, there are several unexplored yet significant research areas. The primary research gaps in AI education that this work addresses are as follows:

1) Though language as a barrier to computing education has been considered by Banerjee et al (2018), most of the research fails to take into account other socio-economic conditions relevant to marginalized children with little or no formal schooling.

2) Studies have shown that children develop a more nuanced understanding of AI when involved with actual code for AI agents and that scaffolding can lead to optimal learning (Druga et al, 2021; Hitchen et al, 2019). However, what we don’t know is how children from low-income backgrounds and lack of access to schools react to this exposure to code and what scaffolding might be required in such demographics.

3) In an in-depth review of resources for AI education by Druga et al (2022), it was found that most of these resources cover only a few of the 5 big ideas of AI4K12 and no existing resource covers all five.

2. Method

2.1 Participants

The participants for this study were out-of-school children in the age group of 9-13. These are children who have once attended school but have since left and don’t plan to rejoin due to financial or social constraints. Though these children had attended schools till different grades, their numerical aptitudes were more or less similar. For example, an 11-year-old boy claimed to have attended school till fifth grade but was unable to perform single digit additions. This can be attributed to poor instruction in previous schools and frequent absenteeism on the student’s part. Therefore, the children in this study were essentially at similar educational levels. The study participants were the children of migrant workers working low-paying jobs like domestic help and manual labour in India. They were selected primarily on the basis of their socioeconomic background. The total number of children who participated in the study were ten and all except two showed basic numeric literacy (the ability to count and recognize numbers from zero to ten) or had some understanding of smartphones (experience using YouTube and Google assistant features). None had any English proficiency and all had Hindi as their native language. Six of the ten children identified as male and the remaining as female. The proposed activity was conducted in a single sitting with all sessions taking place one after another, taking about 50 minutes to complete on average.
2.2 Supporting Learning Resource

The teaching resource used in this study is very simple to develop and can easily be recreated even without much technical proficiency. The autoencoder and the associated interface was developed in the Google Colab, a cloud hosted development environment for python and machine learning. An Autoencoder (Hinton et al, 1989) trained on the MNIST dataset to generate hand-written digits was programmed using TensorFlow. This code, however, is unlikely to be understood by the child participants because 1) it requires proficiency in English and the target demographic for this study is non-English speaking and 2) codes for neural networks are not trivial.

The exact architecture of the autoencoder is not introduced to the participants for the sake of simplicity. This study’s focus is on teaching the broad concept of encoders, decoders and latent space representation. To allow the children to interact and tinker with the code, variables for the latent feature vector (सीखने की क्षमता) and number of epochs (इतनी बार दोहराए) in the code are parameterized with the forms feature in Colab. The value of the latent feature parameter can be adjusted by choosing a value from a dropdown list with three values: 0, 3, 6. Setting it to zero would render the model useless, this can be understood as the neural network not creating any representation of the input and thereby not learning anything. Higher dimensions of this vector means that it is retaining more information to reconstruct images from. Similarly, the parameter for the number of epochs is also presented as a dropdown menu and the training process is visualized with a progress bar. Finally, an option to supply the trained neural network with user input is provided. These elements are illustrated in Figure 2.

2.3 Study Procedure

The study introduced autoencoders to the children in a series of logically segregated sessions aligned with the AI4K12 big ideas (Figure 3). These sessions help structure the learning process, thereby simplifying both the learning and instruction. Also, they can be held over a
course of multiple days which can be used as a scaffolding technique. However, in this study, these sessions were held consecutively on the same day. The 3 sessions are described below:

**Session 1 (Develop analogies):** The session starts with a general introduction of computing devices using the example of smartphones as these children are most likely to have access to these devices. This discussion serves two purposes, it helps the instructor gauge the aptitude of children with respect to computers and clarify any related misconceptions. Children are then introduced to Big Ideas 1 (perception), 2 (representation) and 3 (learning) using analogies like the ones described in Figure 3 (a). The experience of learning in a classroom setting was used because the participants found the scenario relatable.

**Session 2 (Introducing code and interface):** The personal analogies form the basis for developing an intuitive understanding of the functioning of an autoencoder. The first 3 big ideas are sequentially introduced to the children with support from the analogies established in session 1. This progression is apparent in Figure 3 (b). The idea is to help children see the neural network as a learning entity like themselves; this is pretty much like applying the theory of mind (Ensink et al, 2010) for the computer. This process takes place while discussing the Colab notebook containing the code and interface elements. Showing code helps distinguish the AI model as a distinctly human made entity, similar yet very different to humans. Exposure to code also aids the development of a more nuanced understanding of AI (Druga et al, 2021).

The general flow of interaction with the interface is illustrated in Figure 2. Initially the child uses the run button to run a cell containing a piece of code, completion of its execution is reflected by a green tick mark. Then, the child is encouraged to experiment with the model by altering the given parameters in the interface. Once a familiarity is established, the “show code” option is used to reveal the corresponding code to the child. This procedure can repeat itself multiple times depending upon the number of cells and exposed parameters in the notebook. In addition to this, the children are allowed to test the performance of the model on custom inputs for different latent space sizes.

**Session 3 (Usability and impact):** AI4K12’s fourth (natural interaction) and fifth (societal impact) big ideas are discussed in the final session. First goal of this session is to elicit a discussion on how the deep learning model is intelligent but unusable for any practical application involving other people. Second, a more open-ended discussion on the potential usage of such models for social good caters to the requirements of the fifth big idea. This discussion naturally segues into potential biases of such models.

3. Analysis and Findings

The evaluation is done based on perceptions observed during one-one sessions and interviews. Efficacy of instruction was gauged based on the child’s ability to understand and relate to the presented analogies, their understanding of the presented neural network and its visible parameters, comprehend the associated social implications, and suggest novel use cases for the AI model. A qualitative analysis led to the following key findings.

**A simple form-like interface can improve tinkerability and inquiry of neural networks:** Since the children recruited for this study were not fluent in English, they found it difficult to tinker with the program (like by changing the number of epochs) as they couldn’t read through the code to find the variable to alter. The supporting interface proposed above removes this need to understand the code and reliance on English for tinkering with the neural network. Children in this study could easily find the relevant parameters when provided as interface elements in their native language. Since the options for such parameters were limited by a dropdown menu, children found it easier to draw meaningful insights from their inquiry into the functioning of neural networks. This also prevented them from feeling overwhelmed as the number of options presented was carefully limited such that only valid permutations could be formed.
Figure 3. The overall progression of the proposed activity a) sample analogies developed around the AI4K12 big ideas b) equivalent explanations relating the analogies to the actual functioning of autoencoders c) segregation of the five big ideas into a series of sessions.

**Showing code can elicit a sense of agency:** Children could easily interact with the notebook using the interface elements but this alone was not sufficient to elicit ownership and a thorough understanding. The children were unable to think of the AI agent as a programmable entity when no code was shown. They simply saw it as another black box application which they merely use and not actively contribute to (despite them tinkering with the code parameters). However, when code was shown and it was pointed out how the code changes when they choose values from the dropdown menu, the kids felt more in control. This encouraged them to try new combinations of values and feel more confident in dealing with the deep learning notebook.

**Analogies and discussion can help accommodate most AI education goals:** Though the learning resource presented here did not support all five big ideas, the three sessions proved to be efficacious in incorporating all the five goals. When analogies related to learning were established before introducing the code and the interface, the participants’ understanding of autoencoders seemed robust and well-informed. The clear progression of the three proposed sessions allowed for a smooth transition from one idea to another.

**Instruction would have to take place individually:** During the group instruction, significant challenges were faced that prevented the sessions from being successful. It was difficult to engage all the kids in even smaller groups as they tended to get distracted by their environments. For example, in many instances, the child was called by their parents to help with chores like fetching water or collecting donations. Also, unlike a regular classroom, the children in the group had significantly different aptitudes due to differences in prior experience. These differences made it difficult for some children to focus or effectively participate. Some such children also acted out and prevented others from engaging. However, they were able to comprehend and contribute effectively in individual instruction. Individual instruction also helped deal with the unique needs of these children.

**Basic numerical literacy and experience with modern computers is the bottleneck:** Children who did not possess basic numeracy skills were unable to make sense of the second session. Basic numeric literacy is required to tinker with parameters and make sense of the task performed by the neural network. These kids seemed disengaged even in individual interactions, but were able to understand and meaningfully engage in the first and third sessions. A similar response was observed with children who had no prior experience with computers, they couldn’t grasp the idea of AI as they had no prior experience. Such
children couldn’t proceed past the first session. The participants who met these two criteria sufficiently understood the high-level functioning of autoencoders and were able to generalize their functionality to other generative tasks like reconstructing alphabets or sentences.

4. Conclusions, Limitations and Future Work

After a case study with ten out-of-school children with minimal numeric literacy and no English language proficiency, it was found that AI concepts can be meaningfully understood by such children irrespective of other educational differences. Use of analogy driven discussions was shown to effectively cover all five big ideas described by AI4K12 when instruction took place individually. Despite being unable to understand code, exposure to code made children feel more in control when this exposure was mediated with a simple form driven interface. These interfaces are extremely scalable and can be easily created with little technical expertise, thus allowing teachers or volunteers lacking in appropriate training to effectively be able to use such tools in their instruction. Nonetheless, this study was limited by its short duration and small sample size. Future research could include longer interventions and comparative studies with control groups to strengthen the generalisability of these findings. Methods for effectively discussing the AI4K12 big ideas in greater detail should be pursued while also adapting and refining the existing progression charts and policies to cater to out-of-school children. Perceptions and challenges in translating the proposed approach for teaching neural networks in greater depth and breadth also remain to be explored.

References

Informatics Education for University Students based on Text Input Time

Yuko MURAKAMI* & Tomohiro INAGAKI a
a Information Media Center, Hiroshima University, Japan
*yuk0mura@hiroshima-u.ac.jp

Abstract: Japan is rapidly changing the ICT environment for students due to the Global and Innovation Gateway for All (GIGA School) since 2019. Students educated in various ICT environments will soon enter universities, and higher education will need to be optimized in these environments. For current university students in the transitional stage of ICT education, we measured the time required to input text using multiple methods, including laptops and smartphones, and the age when they started using the devices. For university students, smartphone input time is shorter than laptop input time, and individual differences are insignificant, suggesting that smartphone use may be effective in higher education. We also observed that the students who use devices such as smartphones and laptops from earlier educational stages have better typing skills. It will be necessary to design classes assuming a variety of students.

Keywords: ICT device, Informatics Education, Touch-typing, Smartphone, Laptop

1. Introduction

Information literacy is one of the essential liberal arts required today. To acquire information literacy, using ICT in learning activities is recommended; PISA reported that among OECD member countries, Japan spent the least amount of time using digital devices in school classes (NIER, 2019). The Ministry of Education, Culture, Sports, Science and Technology announced the Global and Innovation Gateway for All (GIGA School) to expand the educational ICT environment, making one terminal per student available nationwide in elementary and junior high schools (MEXT, 2019). This terminal includes not only laptop computers but also tablet devices. They also have established an intra-school communication network for elementary, junior high, and high schools.

With realizing the GIGA School, universities will enroll students who have studied in various ICT environments. One can imagine that new students are particularly susceptible to the informatics education curriculum they have received through high school. Most current informatics education assumes the use of laptops, and only individual practice reports have been published on teaching methods that use other types of terminals. It is considered inappropriate for students to take classes based solely on the laptop, as in the past. Examining how ICT can provide better education to university students is necessary.

However, the ability to use information devices is not apparent regarding current university students, for whom the environment for information education was not developed. Our research question is to what extent they can use information devices. We surveyed the age at which students started using laptops and smartphones, the time required for typing as an ICT skill, and students' willingness to use devices. We also discuss effective informatics education for students who will be educated under the GIGA School in the future.

2. Related research

Japanese university students are known to be familiar with the operation of mobile phones. Teaching materials are also being developed using tablets and cellphones. Delcker, Honal,
et al. (2018) suggested from a survey of U.S. university students that tablet devices may provide new university students with the ability to find effective learning strategies and easy access to course materials. The effectiveness of regular delivery of English teaching materials via cell phones was reported to improve reading and grammar skills at a Japanese university (Smith & Wang, 2013). However, the effectiveness of ICT education in Japan does not appear uniform. Gougeon & Cross (2021) suggested differences among Japanese universities in assessing ICT skills.

3. Method

This study measured the time spent touch-typing on a laptop, typing on a smartphone, reading aloud sentences and writing on paper among first-year university students at University H. There are 1345 eligible students, including 187 in the School of Medicine, 410 in the School of Engineering, 81 in the School of Dentistry, 127 in the School of Biotechnology and Production, 130 in the School of Integrated Arts and Sciences, 54 in the School of Pharmacy, and 201 in the School of Science.

Two types of texts were prepared for input exercises. The texts are shown in romaji input, which is commonly used to enter Japanese on a keyboard,

2. AI toha Artificial intelligence no ryaku desu. Artificial toha ‘jinkou teki na’, intelligence ha ‘tisei’ to iu imi de, jinkou tinou to yakusaremasu.

Sentence 1 is a passage from ‘I Am a Cat’ by Soseki Natsume. It is assumed that students who use Japanese daily can read the text easily. Sentence 2 is a brief explanation of Artificial intelligence. Students are required to input Japanese characters and English words.

The time measurement method was not specified. Each student measured the time in each possible way and entered the assignment form. Some students reported times in units of time finer than seconds. However, all data were corrected to the nearest second in the analysis since the accuracy of the reported time varied with this method.

To investigate the relationship between the timing of device use initiation and text input skills, a questionnaire was sent out on when each device was first used. Five options were available: before elementary school, elementary school, junior high school, high school, and after high school. We also surveyed preferences for the terminal used for text input. The survey target, University H, has adopted a “Bring Your Own Device” policy, requiring students to provide their own laptops for use in class. Therefore, almost everyone surveyed should have a laptop and a smartphone and be able to answer these surveys.

4. Result

Figure 1 lists the time taken by each input method for sentences 1 (left (a)) and 2 (right (b)). Using Tukey’s outlier removal method, data exceeding 1.5 times the interquartile range are removed. The figure displays the results of outlier removal. Notches in the figure mean 95% confidence limits the median. In both cases, sentences 1 and 2, the input time on the smartphone is faster than the input time on the laptop. The variation in input time is found to be smaller for laptop, smartphone, hand, and aloud, in that order. This trend is more pronounced for sentence 2 than for sentence 1.

We compared the average input times for laptops and smartphones. Table 1 shows the mean input time for sentence 1 and 2. In the case of sentence 1, the t-test results showed t = 24.912, df = 2194.5, confirming a significant difference at p<0.001. The same analysis was performed for sentence 2, resulting in t = 13.922, df = 2536.2, also a significant
Figure 1. Text input time for laptops, smartphones, handwriting, and speechreading.

Table 1. Mean of Input time for sentences 1 & 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence 1</th>
<th></th>
<th>Sentence 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Num</td>
<td>Mean (S.D.)</td>
<td>Num</td>
<td>Mean (S.D.)</td>
</tr>
<tr>
<td>Laptop</td>
<td>1</td>
<td>303 (7)</td>
<td>1</td>
<td>312 (9)</td>
</tr>
<tr>
<td>Smartphone</td>
<td>2</td>
<td>275 (49)</td>
<td>2</td>
<td>288 (7)</td>
</tr>
</tbody>
</table>

difference at $p<0.001$. It can be said that students spend more time typing on laptops than on smartphones.

Figure 2 summarizes the mean input time for sentence 1 by the time when the device was first used. The links connecting the elements at the top of the figure are represented by "***" for significant differences at $p<0.01$ and "**" for substantial differences at $p<0.05$ based on multiple analyses by ANOVA. Figure 2(a) summarizes the age of starting to use laptops and input time. It can be seen that those who began using laptops after graduating from high school spent more time (77s, 95%CI is from 74 to 84) inputting data than those who started using laptops before entering elementary school (61s, 95%CI is from 53 to 68), elementary school (70s, 95%CI is from 67 to 72), or junior high school (71s, 95%CI is from 67 to 74). Even if they started using laptops in high school, their input time is found to be longer than that of those who started using laptops before junior high school. If students want to improve
their touch-typing skills, it is adequate to start using laptops from elementary school.

Figure 2(b) summarizes the text input time by the start of smartphone use. None of the subjects were students who began using smartphones before elementary school. It was found that those who started using smartphones in high school (52s, 95%CI is from 42 to 63) took longer to input text than those who started using smartphones in elementary (45s, 95%CI is from 42 to 48) or junior high school (49s, 95%CI is from 47 to 50). As with laptops, starting to use smartphones in elementary and secondary school is more effective in improving smartphone input skills.

Table 2 summarizes the results of a survey of students on which device was faster to input devices for sentence 1 and whether they would prefer to input 200 characters on a smartphone or a laptop. 579 students (about 43% of all) preferred the laptop for inputting 200 characters, but were slower to enter sentences and faster on the smartphone. This indicates that some students' preference for text input devices does not match their choice.

Table 2. Cross table of devices that had the fastest input time for sentence 1 and devices that took the student less time to input 200 characters.

<table>
<thead>
<tr>
<th>Devices with short text input time</th>
<th>The device you want to use to type 200 characters.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Laptop</td>
<td>Smartphone</td>
</tr>
<tr>
<td>Laptop</td>
<td>222</td>
<td>26</td>
</tr>
<tr>
<td>Smartphone</td>
<td>579</td>
<td>492</td>
</tr>
<tr>
<td>Same</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>821</td>
<td>524</td>
</tr>
</tbody>
</table>

5. Discussions

5.1 The Importance of ICT Education in the Elementary and Secondary Curriculum

A comparison of the time spent for each input method showed that keyboard input time to the laptop varied widely among individuals. Furthermore, there was less variation in the time spent on handwriting and reading aloud. The phenomenal speed with which the Japanese write on paper was also noted by Houser & Thornton (2004). This is likely because it is repeated early and integrated into the primary and secondary education curriculum (Ministry of Education, Science and Culture, 1980). This suggests that teaching methods such as handwriting and reading aloud may reduce individual differences in device input skills.

A study of the relationship between the timing of device use and input time showed that students who started using devices by junior high school had a faster input time than those who began using devices after high school age. This result suggests that ICT education would be necessary in primary and secondary education.

5.2 Why Text Input Times Differ Depending on When Device Use Begins

This study found that students using laptops in primary and secondary education had shorter text input times. This can be attributed to the fact that ICT education, such as touch typing, is being provided enthusiastically in some Japanese primary and secondary schools. However, special education was not provided for entering text into smartphones, and this reason cannot explain the shorter time required to enter text on smartphones. For both devices, students who had had the devices for longer periods had shorter text entry times, possibly due to their daily experience with the devices.

Another possible cause is the influence of handwriting skills on device input skills. A study comparing touch-typing and handwriting speeds among elementary school students in
the United Kingdom reported that students with no handwriting speed problems were able to type faster on the keyboard than by hand in the upper elementary grades, even without any education on keyboard input (Connelly, V., Gee, D., & Walsh, E., 2007). Our study results show that students are faster at typing by hand. Therefore, these handwriting skills may have influenced their input time to smartphones and laptops. However, this study cannot reveal the causes of these factors; continued research is needed to clarify the impact of continuing education, especially for students who will enter university in the future as a result of the GIGA School, since their learning process up to that point was different from that of previous students.

5.3 University Information Education: Optimal Content and Environment

Smartphone input is faster than laptops, indicating individual variation. This suggests that smartphones save time in situations such as simple communication in class or a quick exchange of opinions in group work. They are also expected to be more efficient in writing reports and essays. Hubert (2017) discussed using English smartphone applications for Japanese students learning English. Although laptops can also use English translation functions, smartphones are faster to type and more suitable for university students.

Some papers question the inclusion of text input methods in informatics education for university students. Zollman (2012) argued that voice input technology advancements will make typing skills unnecessary, eliminating the need for text input classes. Voice input performance has improved significantly in terms of speed and accuracy, as reported by Ruan (2018). University students were found to read text than they typed it, suggesting that lectures on using voice input for better outcomes align with informatics education goals.

For students who have low typing speeds when they enter university, it is better to avoid situations where slow typing speeds prevent them from participating in classes rather than to increase their typing speeds. They are generally given remedial courses, but it is better to teach them practical ways to output what they want to write as quickly as possible rather than just having them do the typing. For example, voice input is suggested for students who take a long time to type on smartphones and computers. In this study, as reading aloud takes the shortest time. Using a voice input system is expected to encourage students to participate in the class actively.

Around 40% of respondents prefer labor-intensive text input devices. This may a lack of accurate understanding of ICT utilization skills among students. This highlights the need to incorporate opportunities for students to gain correct knowledge of ICT utilization skills into their learning. A more detailed survey of student attitudes is necessary to clarify the relationship between text input device choices and self-assessment of ICT utilization skills.

It is increasingly difficult to specify which devices to utilize for university students. ICT education is now provided from elementary education, and many students will have some device before their university course. Thus, it isn’t easy to strongly request university students to purchase a specific device for their university classes. In particular, classes such as information literacy are held in the first year. Their content is aimed at learning a wide range of primary education, so there should be no discrimination in the ease of learning depending on the terminal students’ own. Teachers should develop classes assuming students can access class materials and practice from various devices.

6. Conclusion

This study aimed to clarify the extent to which current university students have the ability using information devices and to examine how information education for university students should be conducted. To this end, we investigated text input time by device and the age at which students began using the devices. Since individual differences in text input time on
smartphones are smaller than on laptops, smartphones would be more time-efficient for simple communication in class and for exchanging opinions in a short period in group work. The study also showed that students who owned ICT terminals in elementary and junior high schools spent less time inputting text on laptops and smartphones. Therefore, when the children and students who are the target of the Global and Innovation Gateway for All (GIGA School) become university students, their text input skills may have improved more than those of today’s university students. The devices they use in elementary and junior high schools and high schools vary, as like laptops or tablets and so on. It is difficult to identify which devices they have the best input skills. Individual students may have different input methods in which they excel. It isn't easy to decide which device is more suitable for them. University classes need to construct course content that does not limit the types of ICT devices.

Acknowledgements

We would like to thank Naoki Yaegashi and Yusuke Namekawa for the discussions. TI was supported by JSPS KAKENHI Grant Numbers JP23H01007 and JP23K11349.

References


Investigating Student Teachers’ Learning Experience and Choice of Participation Modalities in a HyFlex Course: A Mixed Methods Approach

Liang Jing TEHa, Su Luan WONGb*, Mohd Zariat ABDUL RANIc, Mas Nida MD KHAMBARD, & Sai Hong TANGe

a Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia
b Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia
c Faculty of Modern Language and Communication, Universiti Putra Malaysia, Malaysia
d Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia
e Faculty of Engineering, Universiti Putra Malaysia, Malaysia
*suluan@upm.edu.my

Abstract: In this paper, the authors adopted a mixed methods explanatory sequential approach to examine student teachers’ satisfaction and its relationship with participation modalities in a HyFlex course. It also explored the factors behind their extent of satisfaction and choice of participation modalities. Twenty-six student teachers completed the satisfaction questionnaire, and four participants were recruited for a focus group discussion. The research findings indicated that students were very satisfied with the HyFlex course because the flexibility in attendance and the easy access to learning materials accommodated their learning needs and personal commitments; this promoted self-directed learning and reduced their academic pressure. The student teachers were also satisfied with the HyFlex course because they gained exposure to an innovative learning format. As for the factors affecting student choice of participation modalities in the HyFlex course, peer influence is one of the significant factors. Students also reported that they chose the synchronous mode based on their learning preferences and empathy towards the instructor’s commitment, while they chose the online options due to their energy level, health conditions, and personal matters. Overall, this study suggests that HyFlex may be considered as a learning model that could cater to various learning needs among student teachers majoring in baccalaureate in education (physical education), and potentially enhance their learning outcomes.

Keywords: HyFlex, satisfaction, participation choice, student teacher, blended

1. Introduction

Student dropouts have become pervasive in Asia (Latif et al., 2015). Since the coronavirus disease 2019 (COVID-19) outbreak, Malaysia has also been no stranger to this problem. The New Straits Times (2022) and FMT Reporters (2022) reported over 17,000 dropouts among public university students in 2021, which marked an increase of over 4,000 compared to the previous year. In contrast, more than 20,000 students postponed their studies for private institutions. During the same period, Malaysia also suffered a woeful learning loss rate of 0.95 years, one of the highest learning losses among Asian developing countries (Ferlito et al., 2021). Although COVID-19 has brought unprecedented obstacles to education, it boosted students’ adoption of online learning three-fold (Robert, 2022), allowing institutions to propagate innovative and flexible instructional approaches. Hybrid-Flexible (HyFlex) course is one of the learning models that flourished as a remedy in mitigating the challenges imposed...
by the closure of institutions due to the pandemic (Verrecchia & McGlinchey, 2021; Harris et al., 2020) although the concept of HyFlex has existed for more than a decade (Beatty, 2019).

A HyFlex course “provides the capacity for programs to serve remote students in addition to providing convenience and alternatives to regional students” (Beatty, 2019, p. 12). It allows student to choose from one of the three participation modalities, namely physical (attend physical classroom), online synchronous (attend via online conference call platform), and online asynchronous (watch the recording of the class session). Students can change their participation for each class session depending on their personal preferences or needs (Beatty, 2019). The HyFlex learning model offers various benefits such as increasing learning access and flexibility to the learners (Lakhal et al., 2017), providing enriched learning materials that cater to diverse learning styles and preferences (Abdelmalak, 2014), enhancing students' learning outcomes and course completion rates (Irvine et al., 2013), as well as enabling institutions to have higher student enrolment and revenue, reduce unnecessary cost, and create crisis-proof courses (Beatty, 2019). However, implementing this model poses challenges like additional workload for the instructors (Lakhal et al., 2017), managing engagement among both online and in-person learners (Conklinia et al., 2017), and technological issues (Bower et al., 2015), which can affect students' learning experience. Hence, this study aims to investigate learning satisfaction and choice of participation modalities as well as the factors affecting both variables among student teachers in Malaysia.

2. Literature Review

There are several studies which evaluated student satisfaction towards HyFlex courses. A qualitative study was conducted by Abdelmalak (2014) which explored the satisfaction of six university students who joined a HyFlex course and found that students have more control over their learning. Their schedule, learning styles, and needs were accommodated by the flexibility of participation offered by the HyFlex course. Lakhal et al. (2014) did a quantitative research with 439 undergraduate students enrolled in a 10-week HyFlex course, the study revealed that satisfaction level among synchronous students was significantly higher than the students who mostly attended the course asynchronously. A mixed methods study with 311 tertiary level students done by Gobeil-Proulx (2019) found that the flexibility and convenience provided in a HyFlex course were greatly appreciated by students. Interestingly, about 70% of the students stick to one type of participation modality throughout the semester, but the reasons behind this phenomenon were not explained. Rhoads (2020) also did a mixed method study on a HyFlex learning format which comprised 81 undergraduate students; the qualitative findings showed that student satisfaction had a positive relationship with flexibility in attendance. Similar results regarding satisfaction towards the HyFlex learning format were also found by Kakeshita (2021), who showed that around 85% of the 135 students reported high satisfaction with the HyFlex course because their learning needs were met, findings from the article also indicated that students revisit the learning content as recordings of the lecture were given.

Nevertheless, most of the aforementioned studies also revealed a variety of predicaments associated with the HyFlex model. Students reported difficulties staying engaged in the online modalities and were concerned about the quality and consistency of instruction across different modalities (Gobeil-Proulx, 2019). It was also found that some students still prefer the traditional learning mode as it motivates them to complete their learning tasks, and students felt it easier to engage with their instructor should they have any questions (Rhoads, 2020). Furthermore, Liu and Rodriguez (2019) mentioned that students needed help understanding the HyFlex learning format, which caused them to underutilize the learning resources and have less interaction with the instructor.

Compared to the number of studies on online teaching, “research regarding students' perceptions of HyFlex is limited” (Abdelmalak & Parra 2016, p.20), particularly on the factors affecting students' choice of participation modalities in a HyFlex course. Students' attendance
influences learning outcomes, but the impact may vary depending on the teaching modality and individual learning styles (Green, 2021). Students' decisions on the participation mode can be influenced by logistical factors, health conditions, and personal obligations (Blankson et al., 2014). The Unified Theory of Acceptance and Use of Technology (UTAUT) model can explain why students chose a particular participation mode. The UTAUT model is a theoretical framework developed to understand and predict individuals' acceptance and use of technology. According to Venkatesh et al. (2003), four key factors that would influence one's behavioral intentions and adoption of technology were identified, namely performance expectancy (users' belief that using the technology will enhance their performance and productivity), effort expectancy (users' perception of the ease of use and the effort required to use the technology), social influence (the impact of social factors, such as social norms and peer influence, on users' acceptance and use of technology), and facilitating conditions (the extent to which users perceive that the necessary resources and support are available to use the technology effectively). To put these factors in the context of participation modalities selection among the students in a HyFlex course, an introverted student may choose to participate via online while more extroverted students may choose the physical mode (performance expectancy), learners that are staying further away from campus or have a packed schedule might opt for the online options (effort expectancy), one might follow when their friends decided to participate via a specific learning mode (social influence), and students that have connectivity issues or are less familiar with online learning tools may lean towards the physical mode of learning (facilitating conditions).

Given the mixed perceptions towards a HyFlex course among the students, as well as the lack of studies regarding HyFlex learning, there is a need to investigate students' learning satisfaction and the reasons behind their decision of participation modalities in a HyFlex course. As such, the research questions addressed in this study are the following:

1. What is the extent of student teachers' satisfaction with the HyFlex course?
2. Is there a significant relationship between participation modalities and students' satisfaction in a HyFlex course?
3. What factors affect student teachers' satisfaction with the HyFlex course?
4. What factors affect student teachers' choice of participation modalities in a HyFlex course?

3. Methodology

3.1 Research Design

This study employed an explanatory sequential mixed methods approach. According to Creswell and Clark (2018), this research approach begins by conducting the quantitative phase. A subsequent qualitative phase follows it to dive deeper into specific results from the initial phase. The main intent of this design is to utilize the qualitative component to explain the quantitative results.

3.2 Participants and the HyFlex Course

A total of 56 student teachers from the Bachelor of Education (Physical Education) programme registered for the educational technology course which was taught for 14 weeks at the end of 2022. Three participation modalities were provided to all students. For each class session, they were allowed to choose whether they would like to participate by being physically in the Putra Future Classroom (PFC) (in-person), via Zoom or Skype (online synchronous), or by watching the recorded lectures (online asynchronous). PFC is equipped with overhead cameras, high-speed broadband, multiple TV screens, and wireless audio system, allowing
face-to-face and online synchronous students to interact. The recordings of the weekly lecture and the learning materials were uploaded to the learning management system accessible by the students.

For the quantitative phase of this study, 26 students (14 males, 12 females) completed the questionnaire, four participants (2 males, 2 females) from the same group were purposively recruited to participate in the focus group discussion (FGD), two (1 male, 1 female) of the four participants mostly participated the HyFlex course physically while the other two (1 male, 1 female) mostly participated the course online synchronously.

3.3 Data Collection and Data Analysis

A satisfaction questionnaire which comprises 11 items were adapted from Shek et al. (2022) to measure student teachers’ satisfaction towards the HyFlex course. Shek et al. (2022) tested the reliability of the items in 11 different lectures and obtained an average Cronbach’s alpha value of more than .97. In this study, the learning satisfaction measured was segregated into two categories, namely course design (6 items) and instructor attributes (5 items). All the items used a five-point Likert format, ranging from 1 to 5 (1 = strongly disagree to 5 = strongly agree). Table 1 illustrates the 2-times-2 contingency table of participation modalities and satisfaction. Based on the table, the assumption of chi-square test is not met as 50% (more than 20%) of the cells have expected count less than 5. Thus, the Fisher’s Exact test was used, this test is also particularly useful for a small sample size and a two-times-two contingency table (Sprent, 2011), which fits this study well.

After quantitative data analysis, a FGD was conducted to follow up on quantitative findings requiring more context. Qualitative data were collected via a combination of structured and unstructured questions. Before the FGD, prompt questions were pre-prepared according to the research objectives and throughout the FGD, open-ended questions were also asked spontaneously so that the participants’ responses are not limited to the researcher’s bias (Creswell & Creswell, 2018). For qualitative data analysis, content analysis is employed as it allows researchers to systematically identify key themes within the FGD transcript (Erlingsson & Brysiewicz, 2017). All four students’ names were modified to S1 to S4 in the qualitative findings section to keep their personal information confidential.

Table 1. Crosstabulation between participation modalities and satisfaction.

<table>
<thead>
<tr>
<th>Participation Modalities</th>
<th>Satisfied</th>
<th>Very satisfied</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly at PFC</td>
<td>8</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Expected Count</td>
<td>7.6</td>
<td>10.4</td>
<td>18.0</td>
</tr>
<tr>
<td>Mostly via Zoom or Skype</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Expected Count</td>
<td>3.4</td>
<td>4.6</td>
<td>8.0</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td>Expected Count</td>
<td>11.0</td>
<td>15.0</td>
<td>26.0</td>
</tr>
</tbody>
</table>

4. Quantitative Findings

As shown in Table 2, most of the respondents (69.2%) regularly participated the lecture physically, followed by online synchronous mode (30.8%), and no respondent reported that
they often participate the lecture online asynchronously. Based on Table 3, more than 80% of the students strongly agreed or agreed that they were satisfied with both the course design and instructor attribute domain of the HyFlex course. The overall mean score for satisfaction towards course design and instructor attribute were 4.52 (SD = .70) and 4.64 (SD = .62) respectively, showing a high rating of overall satisfaction towards the HyFlex course. The Fisher Exact significant value (2-sided) obtained is P = .08, as the significant value is more than the alpha value ($\alpha = .05$), this indicates that no significant association was found between participation modalities and satisfaction.

Table 2. Frequency and percentage of student teachers’ participation modalities preference.

<table>
<thead>
<tr>
<th>Participation modalities</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostly at PFC</td>
<td>18</td>
<td>69.2</td>
</tr>
<tr>
<td>Mostly via Zoom</td>
<td>8</td>
<td>30.8</td>
</tr>
<tr>
<td>Mostly watch recording</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>26</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>

Table 3. Frequency and percentage of student teachers’ satisfaction.

<table>
<thead>
<tr>
<th>Items</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f</td>
<td>%</td>
<td>f</td>
<td>%</td>
<td>f</td>
</tr>
<tr>
<td>1. Overall, I like the learning experience in this HyFlex course.</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>7.7</td>
<td>3</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>2. The course design is well thought out.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>3. It is easy for me to collaborate with my peers in this course.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>3.8</td>
<td>1</td>
<td>3.8</td>
<td>1</td>
</tr>
<tr>
<td>4. This HyFlex course allows me to learn the lecture according to my situation.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>5. This HyFlex course allows me to learn the lecture efficiently.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>6. This course helped me in understanding the concept of educational technology.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>7. The instructor has a good mastery of the learning materials.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>8. The instructor used different methods to promote learning.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>9. The instructor is able to help students understand the topics covered in the lecture.</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.8</td>
<td>0</td>
</tr>
</tbody>
</table>
5. Qualitative Findings

The factors behind students’ satisfaction were explored in the qualitative phase. This section begins with reasons why students were satisfied with the HyFlex course, which comprises (1) meeting learner’s needs, (2) promoting self-directed learning, and (3) providing affective and professional benefits. It is followed by the factors affecting student choice of participation modalities in the HyFlex course. It was reported that students chose the synchronous mode based on their personal preference and empathy towards the instructor’s commitment while they chose the online options due to energy level, together with health and personal matters. Peer influence plays a significant role in affecting their decision too.

5.1 The Flexibility of Attendance Fulfils Learners’ Needs

When asked about what they liked about the HyFlex course, one of the students stated that the flexibility provided by HyFlex is what he desired most, “…we can choose which [participation modalities] we want or the one that is most suitable for us... If we can’t attend live, we can still watch back the recording” (S1). Another student recalled one scenario during that course and shared: “…I remember why I joined via online is because I was tested positive for COVID, and during that time everyone was doing group discussion...” (S2), that student continued sharing by holding up her phone with one hand and waved towards her phone with another hand: “…and they video called me from the physical classroom like ‘hi’” (S2). The same student shared the struggles faced by the students in attending lectures: “although time is allocated for each subject, most of us have commitments like training as we are athletes right, so we would have training outside of campus...” (S2). Other students further added that learning still occurs in spite of the participation modalities which made learning more convenient for the students, “…like in Zoom, we can be divided into breakout rooms and have activities...” (S3). “I agree... the options give us convenience and ease because we can choose the time we learn” (S4).

5.2 The Accessibility of Learning Resources Empowers Self-directed Learning

Another theme that emerged from the participants’ positive experience with the HyFlex course is that students take ownership of their learning process. A student mentioned that her self-learning was enriched by the accessible learning materials provided in a HyFlex course: “it can become a self-learning material too, like we did join the class physically, then after the class, instead of opening the slides and take notes only, we can actually watch back what has been said by Prof. via the recording, so from there we can revise” (S2).

Another student resonated with the fact that students have more autonomy on how they learn in a HyFlex course: “Like for me honestly, I’m the last-minute kind of student, so I would use all the learning resources provided by Prof. and study before the final exam. So it’s very helpful for me because I just click, and all the lectures and learning resources from week one to week 13 are there” (S3). The same student also pointed out that the learning resources provide student with different learning progress the same learning opportunities, making learning more equitable to all students: “because it’s like unfair to those that couldn’t make it to class, as they would not learn the knowledge of that session. Like ‘S2’ mentioned just now,
we still have the learning materials and recordings provided, so there’s no excuse to not study” (S3).

5.3 HyFlex Course Provide Students with Affective and Professional Benefits

When prompted deeper on how the flexibility and autonomy benefits the students, one of the participants expressed that HyFlex learning format reduced her psychological distress: “For me, it reduces my pressure, as in we don’t know right sometimes, we have emergency or other matters that are critical, so I can’t join the class physically, it’s okay because I can have another choice to gain the knowledge that Prof. taught on that session” (S2). Another student agreed by saying: “…we also have back-to-back classes, the distance between two classes can be very far away too, so it will be hard. Like we need to make sure the classes end on time right, let’s say this lecturer does not finish class on time, it will drag and affect your next class. So I think this HyFlex can reduce students’ pressure as they can choose” (S1).

Besides reducing academic pressure, the students stated that HyFlex course can increase their motivation to learn due to the ubiquitous learning resources given: “let’s say the course is only physical, once you missed a session, you will need to approach your friends or the lecturer to catch back on the learning content for that session. When we have HyFlex, like we said before all the lectures and learning materials are there. So you don’t have to like WhatsApp your friend, and your friend would need effort to recall and reply, maybe that friend know the learning content but it’s not 100% accurate like in class” (S3). Intriguingly, a participant perceived being in a HyFlex course as a chance to be exposed to different instructional methods that can potentially enhance his future career, “…we can get a taste of running physical, online synchronous, and online asynchronous class. I can observe how Prof. prepare and conduct the different medium of learning so that all students are not left out in learning. So I think experience like this is important for us especially the ones that are majoring in education for our future career” (S1).

5.4 Factors Affecting Student Choice of Participation Modalities in a HyFlex Course

Regarding the reasons behind students’ choice of participation mode for each class session, the students shared that they chose to participate physically because of their empathy towards the instructor, “…only a few students in physical class and online synchronous, I feel like it's not respecting the lecturer…and I empathize with lecturers that have put in the commitment to teach” (S1). This comment was seconded by another student, saying: “Like what ‘S1’ said, we respect Prof. as our lecturer. How would she feel if she sees four physical students and nine online students only. At that moment she did not show her feelings, but we need to understand, because we are going to be in her shoe also one day” (S2). Students decide their participation mode based on their learning style too, one participant shared at length about this factor: “Like for me, I don’t really like watching recording, some more the videos are always like long…more than an hour… I need to just sit down, it’s really not my style” (S1).

Moreover, students also shared that they chose the online options when their degree of fatigue is higher, “Like even at night time we still have classes and it’s all physical classes, it’s very tiring, and the next morning we also have classes…so it becomes like ahh just watch the recording would be enough la” (S1), “I also have training every day, so will feel like lazy haha, so my option is online. But it depends on the student himself, if [I] feel like [I] have the responsibility to study, then I will go [physical class]. But there are times when the students don’t join even join through online, didn’t join at all…” (S4). When participants were asked about why those students did not join, one participant claimed that: “maybe they are not interested with the subject” (S3), but another participant disagreed and stated that they did not join because of the lack of interest in study as a whole instead of that subject only, and further described the unique learning style of physical education students: “…we are more interested in workshops, do activities instead of sit and listen, can’t focus on one thing for a long time” (S2).
Peer influence was also one of the driving motives that affected student's decision, one student explained how this happened: “Like for example our class is until 11pm at night, so we become really tired right, so when our friends are tired and do not feel like attending, we would follow, and sometimes we would be the ones initiating and them follow too” (S3). Other than these factors, the previous section also indicated that students chose the online options due to health matters like the contraction of COVID-19 as well as personal commitments such as trainings.

6. Discussion

The primary aim of this study is to examine student teachers’ extent of satisfaction towards a HyFlex course and the association between participation modalities and satisfaction, as well as exploring the factors affecting their level of satisfaction and decision on participation modalities in a HyFlex course. It is important to mention that there is a possibility that the students who participated mostly via online asynchronous mode did not respond to the survey as the participant indicated that they do not have an interest in studying in general.

The high overall mean scores for satisfaction towards both the course design and instructor attributes of the HyFlex course found in the quantitative phase were validated by the qualitative findings, the student teachers appreciated the flexibility in participating the class for each week as they were allowed to learn according to their learning preferences and personal schedules. When the student teachers could not attend the class synchronously due to personal commitments or unforeseen reasons like health issues, HyFlex enabled them to learn using their most effective ways and keep up with the learning progress. The flexibility offered by HyFlex was particularly important to the student teachers in this study because the majority of them were athletes, and they had commitments like trainings and competitions that often clashed with the timing of the class sessions. This flexibility increased the ease of learning within the students and helps them balance their academic progress and personal responsibilities, contributing to their overall satisfaction towards the course. This is in line with the findings in the studies done by Rhoads (2020), Gobeil-Proulx (2019), and Lakhal et al. (2017) which concluded that students valued the flexibility and convenience provided in a HyFlex course.

The statistical findings also showed that the level of satisfaction towards the HyFlex course among the student teachers is independent of the types of participation mode, and this can be explained via the responses from the FGD which revealed that students had equal learning opportunities despite their choice of participation mode. This is because the enriched learning materials such as recordings of lectures, reading materials and discussion forums were ubiquitous and accessible to all students within the HyFlex course. This prevented the student teachers from being left out as they could still participate in the lecture asynchronously when they were unable to be present during the class sessions due to personal obligations. This is especially beneficial for the participants who were athletes in this study, as their regular commitment to trainings and competitions would often leave them feeling exhausted, making them more prone to being absent or distracted in class. By providing multiple modes of access to the learning content, all students with different learning needs can have an equal opportunity to succeed in the course. This finding resonated with the research by Abdelmalak (2014) and Kakeshita (2021) which stated that HyFlex courses can cater to different learning needs.

Findings from this study also found that with the flexibility, accessibility, and convenience given in a HyFlex course, learners become not only more satisfied, but also more motivated to learn, making their learning more self-directed (Kakeshita, 2021). This can empower learners to take ownership of their learning by deciding when and how they want to learn. For instance, the responses from the FGD in this study showed that before exams, students would re-experience the lecture via the recordings to review the course materials multiple times for a better understanding. This can potentially improve students’ learning outcomes, reduce learning loss and dropout rates, and increase the overall course completion rate amongst the students (Beatty, 2019; Irvine et al., 2013). This finding is particularly pertinent to Asian learners as their classrooms tend to be more teacher-centered, but
classrooms with a more student-centered instruction within East Asian societies have been found to have higher achievement (Kim, 2018). To the researchers’ surprise, a HyFlex course can also develop students’ future career prospects as they are exposed to the innovative instructional approaches of HyFlex throughout the course. They could learn effective ways of conducting physical, online synchronous, and online asynchronous instruction, which could contribute to their professional growth, considering their future roles as educators. This hands-on experience could also concretize students’ learning about educational technology. This finding is rarely seen in the literature on HyFlex learning, but the researchers acknowledge that this finding would most likely only ring true in cases where the learners of the HyFlex course are majoring in education-related programmes of study.

Furthermore, this study also revealed that although HyFlex can cater to various learning needs, practitioners must be aware of students’ learning styles and tailor their instruction accordingly to ensure the quality of the learning experience. As most of the student teachers in this study were athletes, they admitted that their learning style leans towards the kinesthetics’ side of learning style. Thus, the instructor needs to utilize teaching methods which are more tactile to keep the learners engaged. However, this study indicates that when it comes to student choice of participation modalities, there were factors which are out of the locus of control of the instructor. Student’s decision to participate in a class session can directly influence their peers’ decisions as learners that are close friends tend to stick together throughout the instruction. Findings from this study also suggested that students attended the class synchronously because they felt bad for the instructor if only a few students were present during the lecture. This shows that Asian students may conform to classroom norms due to their respect for authority figures such as instructors in a HyFlex learning environment (Li & Campbell, 2008). Additionally, responses from the FGD suggested that when students are not in good health condition, feel exhausted due to a packed academic timetable and personal commitments like training, or are bound by personal obligations like a family plan which clashes with the lecture’s timing, they would opt for the online options. This finding parallels the study by Blankson et al. (2014), which concluded that student choice of participation mode was influenced by their health, personal commitments, and commuting challenges. In other words, different groups of students would have a different pattern of participation modalities selection as their academic schedule as well as their attitudes towards learning and the instructor would be different. Also, students’ choice of participation mode would be affected during the period of homecoming festivals as students would prioritise travelling plans, or during the onset of the cold and flu season when more learners are more prone to getting ill.

7. Conclusion

As a whole, this study showed that HyFlex course can potentially be a suitable learning model for Asian student teachers who are majoring in bachelor’s degree in education (physical education) mainly because the flexibility in classroom participation and the equity in learning opportunities. The HyFlex learning model was able to accommodate students’ learning needs due to unforeseen circumstances and personal commitments, which enabled learners to have the best of both worlds between their academic progress and personal lives. With the accessible and ubiquitous learning resources, learning becomes more self-directed and more equitable among students, this may alleviate their anxiety about falling behind in learning and allow them to feel more motivated to learn. This could potentially enhance their learning performance as well as reduce dropout rate and learning loss among the students. Other than that, student teachers also benefitted from the HyFlex course in terms professional development as they gained exposure to integrating an innovative learning format. Suffice to say, policy makers, faculties and instructors should consider incorporating the HyFlex learning model into the teaching and learning experience.

Regarding student choice on participation modalities, this study found that students choose their participation mode depending on their empathy towards the instructor, learning style, level of fatigue, and influence from their peers. The trend of students’ choice of participation can be predicted based on factors such as the holiday season and times when
illnesses are more prevalent — leading to a higher number of students choosing the online options. Future studies can look into learning interest or motivation among students and how it relates to their preferred participation modalities in a HyFlex course as this study suggests that students’ choice of learning mode was affected by their learning desire. While qualitative findings from this study have shown that a HyFlex course can promote learning equity, future studies can provide more empirical results regarding learning equity across all participation modalities.

8. Limitations of the Study

A few limitations exist within this study. Firstly, the small sample size which did not encompass students who mostly participated in the lectures asynchronously, and the fact that all participants in this study were from the physical education programme affect the representation of this study towards other student groups, a larger number of samples from a more diverse demographics can be recruited to expand the generalizability of the findings further. Moreover, the data collected were self-reported and were based on the student teachers’ honesty and perception of their learning experience in the HyFlex course.

References


Kakeshita, T. (2021, April). Improved HyFlex Course Design Utilizing Live Online and On-demand Courses. In CSEDU (2) (pp. 104-113).

Kim, Y. (2018). Revisiting classroom practices in East Asian Countries: Examination of within-country variations and effects of classroom instruction. Teachers College Record, 120(7), 1-42.


Assessment of Intelligent Teaching Preparation of EFL Teachers: Based on Two-Year Data Comparison

Xin AN*, Xi SHEN*, Jiannan BAIa & Yushun LIa
*a school of Educational Technology, Beijing Normal University, Beijing, China
* 17888829110@163.com

Abstract: In order to guide the effective application of artificial intelligence (AI) in education (AIEd), China has set up a number of national AI education demonstration areas (AIEDA). Teachers in this AIEDA have carried out action research about AIEd for over two years. For the purpose of assessing teachers’ intelligent teaching preparation during the process, we conducted consecutive two years of evaluation through questionnaires and AI-based teaching platform data. This study developed a questionnaire of teachers’ intelligent teaching preparation, including teachers’ perception of and behavioral intention to AIEd, AI technological pedagogical and content knowledge (TPACK), and AI ethics knowledge. Besides, the data obtained from an AI-based teaching platform were analyzed, which indicated teachers’ acceptance and trust towards AIEd. A total of 101 valid samples of English as a foreign language (EFL) teachers participated in data collection for two years. The results showed that teachers’ trust in AIEd increased, but their effort expectancy and perception of facilitating conditions of AIEd decreased; Teachers’ AI ethics knowledge increased, while their self-report TPACK decreased.

Keywords: Artificial intelligence, EFL teacher, intelligent teaching preparation, AI education demonstration area

1. Introduction

Artificial intelligence (AI) is shaping the education sector and has become an important area of research across the world (Khan et al., 2021). Various countries such as the United States (National Science and Technology Council, 2016), the United Kingdom (Department for Science, Innovation and Technology et al., 2017), France (Government Francais, 2017), and Japan (Headquarters for Japanese Economic Revitalization, 2016), are deploying AI technologies in education. China has designated AI education demonstration areas (AIEDA) to explore the use of AI for subject learning (An et al., 2022), and promote the development of teachers’ preparation of AI in education (AIEd) (The Teachers’ Office of China, 2018). Teachers in K-12 schools in the AIEDA have leveraged AI affordances through multiple AI tools to facilitate personalized learning and immediate feedback since 2020 (An et al., 2022). The integration of AI into education has presented new challenges to teachers, it is crucial to assess the preparedness of teachers in AI environment.

However, existing research on the AIEd environment has primarily focused on students, while research on teachers remains significantly lacking (Liang et al., 2021). Teachers need to be prepared to embrace the changes brought about by AI and work towards developing their own skills and knowledge to meet the demands of the modern-day classroom. Assessing the preparedness of teachers in AI environment can provide them with targeted suggestions for improvement. Therefore, teachers’ intelligent teaching preparations was studied in this research to fill the research gap. Intelligent teaching refers to the application of AI technology and algorithms by teachers to optimize the teaching and learning process and improve students’ learning outcomes (Celik, 2023; Liu, 2020).

This study aimed to evaluate teachers’ preparation to incorporate AIEd by examining both external and internal factors (An et al., 2022). The Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003) accounts for the perceptions of technology from users’ viewpoint and was therefore used to explore teachers’ perceptions and acceptance of AIEd as external factors. The
intelligent Technological Pedagogical and Content Knowledge (iTPACK) framework (An et al., 2022; Celik, 2023) was used to assess teachers' knowledge of AIEd as internal factors. The study conducted a two-year longitudinal survey of the same group of teachers and collected data from both questionnaires and an AI teaching platform to provide a more comprehensive assessment.

2. Conceptual framework

2.1 UTAUT

UTAUT is a widely used model for understanding users’ acceptance and usage of new technologies. It was proposed by integrating the Technology Acceptance Model (TAM) and seven other models. It has been shown to explain up to 70% of the variance in predicting users’ behavioral intention to use new technology (Venkatesh et al., 2003). UTAUT surveys various factors, including performance expectation, effort expectation, social influence, facilitating condition, users’ behavioral intention, and behavior. The definitions of these factors are shown in Table 2. Due to its effectiveness, UTAUT has been widely used in education field (Hsu, 2022). As AI technology is increasingly used in education, UTAUT could be applied to assess students’ acceptance of AIEd. For instance, An et al. (2022) developed a questionnaire based on UTAUT, and explored teachers’ behavioral intention to use AIEd systems and its related factors. However, there is still a lack of annual follow-up surveys on teachers, making it difficult to demonstrate the changes in teachers’ perceptions and acceptance under the rapid development of AI technology. In this study, UTAUT was adopted to explore teachers’ perceptions and acceptance of AIEd in two years.

2.2 iTPACK

Technological Pedagogical Content Knowledge (TPACK) is a framework that describes the knowledge that teachers need to effectively integrate technology into their teaching practices. Developed by Mishra and Koehler (2006), TPACK comprises three core areas of knowledge: technological knowledge (TK), pedagogical knowledge (PK), and content knowledge (CK). These three types of knowledge are interrelated and combine to form Pedagogical Content Knowledge (PCK), Technological Pedagogical Knowledge (TPK), Technological Content Knowledge (TCK), and TPACK. TPACK provides a framework for professional development, helping teachers to identify areas where they may need to strengthen their knowledge and skills. By leveraging this framework, teachers can effectively integrate technology into their teaching practices.

When AI entered education, a series of ethical issues have appeared. It is believed that teachers should master AI ethics knowledge as well as TPACK. Celik (2023) proposed iTPACK framework, including teachers’ TPACK and ethics knowledge. Teachers should have a rational understanding of the ethical issues in AIEd, such as privacy, fairness, and responsibility (Feng et al., 2020). In the process of integrating AI into teaching, teachers need to ensure that they consciously abide by social moral norms in information activities, pay attention to privacy and data security, and deal with issues related to information cocoon, data privacy, AI bias and ethical decision-making in a targeted way (Burton, 2017). However, previous studies typically only conducted one-time surveys, making it difficult to present the knowledge development of teachers in AI environments. In this study, we would conduct a follow-up survey on teachers’ knowledge based on iTPACK framework to fill the research gap.

3. Methods

3.1 Participants

The participants of this study are teachers of English as foreign language (EFL) in middle school in Dongcheng District, one of the AIEDA in China. In Dongcheng District, the Information Office of the Education Commission, the Education Science Research Institute, universities, and enterprises are cooperating to promote the integration of AI and teaching among middle school EFL teachers. Teachers
conducted action research to explore the integration of AI and EFL subjects in this district. Each teacher is guided by three interdisciplinary experts, which includes an EFL expert, an educational technology expert, and a scientific research expert, contributing to the development of intelligent teaching preparation among teachers.

To assess the development of teachers’ preparation for AIEd, questionnaires and AI teaching platform (E-listening and speaking, https://www.ets100.com/home/perintro.html) data analysis were used in this study. Teachers can assign a variety of exercises, which are automatically fed back by the AI to students, and the teacher can also give feedback on the platform. A total of 354 EFL teachers’ AI teaching platform data was collected on April 30, 2022, including the users’ data in the period of January 1, 2021 to April 30, 2022. The first questionnaire was conducted from November 3, 2021 to November 9, 2021, and the second questionnaire was conducted from December 5, 2022 to December 14, 2022. A total of 101 EFL teachers who effectively participated in both surveys were selected, while samples with duplicate options for all scale questions were deleted. The basic information of the respondents is shown in Table 1.

Table 1. Demographic Information of Participants

<table>
<thead>
<tr>
<th>Demographic Profile</th>
<th>Classification</th>
<th>Number</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>2</td>
<td>1.98%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>99</td>
<td>98.02%</td>
</tr>
<tr>
<td>Year of Birth</td>
<td>1960-1969</td>
<td>6</td>
<td>5.94%</td>
</tr>
<tr>
<td></td>
<td>1970-1979</td>
<td>36</td>
<td>35.64%</td>
</tr>
<tr>
<td></td>
<td>1980-1989</td>
<td>39</td>
<td>38.61%</td>
</tr>
<tr>
<td></td>
<td>1990-1999</td>
<td>24</td>
<td>23.76%</td>
</tr>
<tr>
<td></td>
<td>1981-1989</td>
<td>4</td>
<td>3.96%</td>
</tr>
<tr>
<td></td>
<td>1990-1999</td>
<td>20</td>
<td>19.80%</td>
</tr>
<tr>
<td></td>
<td>2000-2009</td>
<td>37</td>
<td>36.63%</td>
</tr>
<tr>
<td></td>
<td>2010-2019</td>
<td>30</td>
<td>29.70%</td>
</tr>
<tr>
<td></td>
<td>2020-2021</td>
<td>10</td>
<td>9.90%</td>
</tr>
</tbody>
</table>

Table 2. Structure of Teacher Intelligent Teaching Preparation Assessment Scale

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sub-Dimension</th>
<th>Definitions</th>
<th>Sample item</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTAUT</td>
<td>Performance Expectancy (PE)</td>
<td>Performance Expectancy means the degree to which a teacher believes that using the AI system will help him or her attain gains in job performance (Venkatesh et al., 2003).</td>
<td>AI can help me improve the quality of teaching.</td>
</tr>
<tr>
<td></td>
<td>Effort Expectancy (EE)</td>
<td>Effort Expectancy means the degree of ease associated with the use of an AI system (Venkatesh et al., 2003).</td>
<td>I think AI teaching systems are very simple.</td>
</tr>
</tbody>
</table>

3.2 Instrument

The questionnaire was developed mainly referred to An et al (2022) and Celik (2023), including teachers’ perception of and behavioral intention to AIEd based on UATUT, and teachers’ knowledge preparation of AIEd based on iTPACK framework. Table 2 lists the constructs and item examples of the questionnaire. Five-point Likert scale was adopted in the questionnaire, with 1 representing strongly disagree, 2 representing disagree, 3 representing general, 4 representing agree, and 5 representing strongly agree. Based on the relevant literature and theory development, the scale was reviewed and filled out by two educational technology experts, one staff member of the Education Commission of Dongcheng District, two junior high school English teachers, and modified according to the feedback to ensure the content validity.

Table 2. Structure of Teacher Intelligent Teaching Preparation Assessment Scale
<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facilitating Conditions (FC)</td>
<td>Facilitating Conditions means the degree to which a teacher believes that an organizational and technical infrastructure exists to support the use of the AI system (Venkatesh et al., 2003).</td>
<td>When I need to use AI in teaching, my school will provide help for me.</td>
</tr>
<tr>
<td>Social Influence (SI)</td>
<td>Social Influence means the degree to which an individual perceives that important others believe he or she should use the new system (Venkatesh et al., 2003).</td>
<td>My colleagues think I should use AI to support teaching.</td>
</tr>
<tr>
<td>Behavioral Intention (BI)</td>
<td>Behavioral Intention means the beliefs teachers have about the AI using behavior under consideration (Davis, 1989).</td>
<td>I intend to use AI in teaching in the future.</td>
</tr>
<tr>
<td>iTPACK</td>
<td><strong>AI language technological knowledge (AI-TK)</strong>: AI-TK means knowledge about AI technologies in language fields (Schmid et al., 2020).</td>
<td>I know that speech recognition technology can score pronunciation, fluency, and phonological tone type.</td>
</tr>
<tr>
<td>AI technological pedagogical knowledge (AI-TPK)</td>
<td><strong>AI-TPK</strong> means knowledge of how teaching may be changed as the result of using AI technologies (An et al., 2022).</td>
<td>I know how to use AI tools to meet students’ differentiated learning needs.</td>
</tr>
<tr>
<td>AI technological content knowledge (AI-TCK)</td>
<td><strong>AI-TCK</strong> Knowledge about how AI technology and subject content relate to each other (Schmid et al., 2020).</td>
<td>I know how to use AI to help students learn vocabulary.</td>
</tr>
<tr>
<td>AI ethics</td>
<td>AI ethics means that teachers can consciously abide by social moral norms in AI-supported teaching activities, have a rational understanding of ethics and safety issues in the process of integrating AI into education and teaching, pay attention to the threat of AI to education ethics and safety and properly deal with it (An et al., 2022).</td>
<td>I will pay attention to the privacy of students' data when using AI.</td>
</tr>
</tbody>
</table>

3.3 *Data analysis*

Following data collection, the reliability and structural validity of the survey scale were evaluated. Data analysis includes two parts: Confirmatory Factor Analysis (CFA) and reliability analysis. Reliability analysis of the scale was conducted with SPSS20.0, while CFA was conducted with Mplus 8.3. The standards recommended by Hair et al., (2014) were adopted in the CFA. Accordingly, indices of $\chi^2$/df ($< 5$), Root Mean Square Error of Approximation (RMSEA) ($< 0.10$), Comparative Fit Index (CFI) ($> 0.90$), and the Tucker-Lewis Index (TLI) ($> 0.90$) were utilized to assess the model fit. To examine the convergent validity of the scale, factor loadings ($\lambda$) were employed to calculate Average Variance...
Extracted (AVE) (>0.5) and Construct Reliability (CR) (>0.7). Discriminant validity of the scale was determined by comparing the square root values of AVEs of components with the correlations between components.

In the reliability analysis of the scale, Cronbach's $\alpha$ value was calculated to assess the internal consistency coefficients. The Cronbach's $\alpha$ value of the entire scale and all sub-scales should be above 0.7 (Fornell & Larcker, 1981).

4. Results

4.1 Validity and Reliability

To verify the structural validity of the scale, CFA was carried out. The model fit indices were assessed, with a $\chi^2$/df of 3.001 (< 5.0), RMSEA of 0.091 (< 0.10), CFI of 0.916 (> 0.90), and TLI of 0.901 (> 0.90), indicating an acceptable fit for the scale items. Standardized factor loadings ranged from 0.719 to 0.949, all falling within an appropriate range. Additionally, the AVE values exceeded 0.5, and the CR was higher than 0.7, indicating good convergent validity. Discriminant validity was also confirmed, with the square root values of AVE for each component exceeding the respective inter-component correlations. Furthermore, the reliability coefficient Cronbach's $\alpha$ of subdomains ranged between 0.880 and 0.939, with an overall reliability of 0.952, indicating high internal consistency. In conclusion, the scale has good validity and reliability.

4.2 Acceptance and Trust for AI Teaching Platform

The real teaching behaviors of teachers are analyzed with the help of backend data from AI teaching platform. The two kinds of backend data (number of assign homework and number of check homework) represent two commonly used functions for teachers on the platform. The results are shown in Table 3. In 2022, the average monthly number of teachers using this platform to assign homework has increased significantly compared with 2021, but the proportion of teachers checking homework is declining. This indicates that teachers’ acceptance of the AI teaching platform is increasing, and their trust in the automatic feedback function of the platform is also increasing.

Table 3. The use of an AI teaching platform by EFL teachers in Dongcheng District

<table>
<thead>
<tr>
<th>Classification</th>
<th>Average in 2021</th>
<th>Average in 2022</th>
<th>Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of assign homework (AH)/month</td>
<td>6.29</td>
<td>16.21</td>
<td>157.69%</td>
</tr>
<tr>
<td>Number of Check homework (CH)/month</td>
<td>1.96</td>
<td>2.74</td>
<td>39.97%</td>
</tr>
<tr>
<td>AH/CH (%)</td>
<td>31.18</td>
<td>16.94</td>
<td>-45.68%</td>
</tr>
</tbody>
</table>

4.3 UTAUT of AIEd

In order to learn more about the situation of teachers, this study conducted a questionnaire survey. Samples of scale questions in the questionnaire are normally distributed (range of kurtosis is [-1.441, 1.254], range of skewness is [-0.463, 0.394]), so it is possible to conduct paired sample t-tests for two years of teacher scale data. The scores of various dimensions and the different test results of UTAUT of middle school EFL teachers in 2021 and 2022 are shown in Table 4. There was no significant difference in the three dimensions of teachers' performance expectancy, social influence and behavioral intention in the two-year survey, while the scores of effort expectancy and facilitating conditions in 2022 were significantly lower than those in 2021.

Table 4. Analysis on the difference of UTAUT scores of middle school EFL teachers in 2021 and 2022
The specific scores of iTPACK of EFL teachers in 2021 and 2022 are shown in Table 5. There is no significant change in TK and TCK, but TPK and TPACK are significantly lower in 2022 than in 2021. In the meanwhile, teachers’ knowledge of AI ethics improved significantly in 2022 compared with 2021.

<table>
<thead>
<tr>
<th>Items</th>
<th>M</th>
<th>MD</th>
<th>SD</th>
<th>95% confidence interval lower</th>
<th>95% confidence interval upper</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Expectancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>4.15</td>
<td>0.02</td>
<td>0.59</td>
<td>-0.09</td>
<td>0.14</td>
<td>0.41</td>
<td>0.68</td>
</tr>
<tr>
<td>Performance Expectancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>4.12</td>
<td></td>
<td></td>
<td></td>
<td>2.25</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Effort Expectancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>3.75</td>
<td>0.15</td>
<td>0.67</td>
<td>0.02</td>
<td>0.28</td>
<td>2.24</td>
<td>0.03</td>
</tr>
<tr>
<td>Effort Expectancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>3.60</td>
<td></td>
<td></td>
<td></td>
<td>2.24</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Facilitating Conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>3.74</td>
<td>0.15</td>
<td>0.69</td>
<td>0.02</td>
<td>0.29</td>
<td>2.24</td>
<td>0.03</td>
</tr>
<tr>
<td>Facilitating Conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>3.58</td>
<td></td>
<td></td>
<td></td>
<td>2.24</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Social Influence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>3.63</td>
<td>0.05</td>
<td>0.71</td>
<td>-0.09</td>
<td>0.19</td>
<td>0.67</td>
<td>0.51</td>
</tr>
<tr>
<td>Social Influence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>3.58</td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Behavioral Intention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>4.03</td>
<td>-0.01</td>
<td>0.58</td>
<td>-0.13</td>
<td>0.10</td>
<td>-0.22</td>
<td>0.83</td>
</tr>
<tr>
<td>Behavioral Intention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>4.04</td>
<td></td>
<td></td>
<td></td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4 iTPACK

The specific scores of iTPACK of EFL teachers in 2021 and 2022 are shown in Table 5. There is no significant change in TK and TCK, but TPK and TPACK are significantly lower in 2022 than in 2021. In the meanwhile, teachers’ knowledge of AI ethics improved significantly in 2022 compared with 2021.

Table 5. Analysis on the difference of iTPACK scores of middle school EFL teachers in 2021 and 2022

<table>
<thead>
<tr>
<th>Items</th>
<th>M</th>
<th>MD</th>
<th>SD</th>
<th>95% confidence interval lower</th>
<th>95% confidence interval upper</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TK2021</td>
<td>3.61</td>
<td>0.11</td>
<td>0.78</td>
<td>-0.04</td>
<td>0.27</td>
<td>1.48</td>
<td>0.14</td>
</tr>
<tr>
<td>TK2022</td>
<td>3.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pair 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPK2021</td>
<td>3.62</td>
<td>0.30</td>
<td>0.64</td>
<td>0.17</td>
<td>0.42</td>
<td>4.63</td>
<td>0.00</td>
</tr>
<tr>
<td>TPK2022</td>
<td>3.32</td>
<td></td>
<td></td>
<td></td>
<td>4.63</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Pair 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCK2021</td>
<td>3.70</td>
<td>-0.07</td>
<td>0.82</td>
<td>-0.23</td>
<td>0.09</td>
<td>-0.82</td>
<td>0.41</td>
</tr>
<tr>
<td>TCK2022</td>
<td>3.76</td>
<td></td>
<td></td>
<td></td>
<td>-0.82</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Pair 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPACK2021</td>
<td>3.84</td>
<td>0.22</td>
<td>0.71</td>
<td>0.08</td>
<td>0.36</td>
<td>3.11</td>
<td>0.00</td>
</tr>
<tr>
<td>TPACK2022</td>
<td>3.62</td>
<td></td>
<td></td>
<td></td>
<td>3.11</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Pair 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI Ethics2021</td>
<td>3.60</td>
<td>0.58</td>
<td>0.63</td>
<td>0.46</td>
<td>0.70</td>
<td>9.40</td>
<td>0.00</td>
</tr>
<tr>
<td>AI Ethics2022</td>
<td>4.18</td>
<td></td>
<td></td>
<td></td>
<td>9.40</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

5. Discussion

5.1 Teachers' trust in AI has increased, as has their expectation for the development of technology environment

According to the teachers' behavior data collected from the AI teaching platform, teachers' trust and acceptance of AI are improving. This is the real behaviors of teachers using AI teaching platforms, indicating that teachers have felt the reliable help of AI. However, compared with two years of questionnaire survey results, it can be seen that teachers' perceptions of effort expectancy and facilitating conditions are declining. This indicates that with the deep integration of AI and subject teaching, teachers have diversified and high-level demands in practice, and their expectations and requirements for AI teaching products are increasing, while the support of intelligent teaching software and hardware environment increasingly lags behind teachers' expectations. In the future, we should continue to promote the improvement of AI EFL teaching products, strengthen the support services of technical personnel, and escort the practice of the integration of AI and education for teachers.
5.2 Teachers' self-report AI pedagogical knowledge declined, but AI ethics knowledge improved

From the comparison results of two years, the scores of teachers in AI ethics knowledge increased, while TPK and TPCK of self-evaluation decreased. This indicates that with the development of expert training and teacher action research, teachers have a more prudent attitude towards AI ethics issues, such as the information cocoon, privacy disclosure, decision-making bias, and accountability difficulties. However, the scores of teachers’ self-report AI technological pedagogical knowledge declined. It might be because that EFL teachers have realized their shortcomings in AI-assisted pedagogical knowledge as AI integrates with EFL deeper and deeper. In the follow-up practice, more attention should be paid to teachers' ability to improve AI-supported teaching, so that they can use AI to implement more teaching strategies and carry out personalized and precise teaching.

6. Implications

This study showed the current situation and development of teachers' intelligent teaching preparation in AIEDA in China. This study provides evaluation instruments for teachers' intelligent teaching preparation in the process of AI and subject integration, and the research findings provide a practical reference for the integration of AI and education in K-12 further. At the teacher level, this study provides teachers with tools for self-evaluation and self-reflection, points out the direction for their future professional ability development, and provides a guarantee for their independent and professional action research.

Due to space limitations, this study does not present the evaluation content of qualitative data such as interviews and analysis of subject text data. In addition, promoting the development of students is the fundamental purpose of integrating AI technology into education and teaching. This study lacks the evaluation of teachers from the perspective of students, which needs to be improved continuously in the future.

Acknowledgements

This research was supported by the National Social Science Foundation "Research on Artificial Intelligence Technical Specifications for Minors" (Grant Number 21&ZD328).

References


Using Learning Design Technologies for Teachers’ Practice-Driven Research

Marc BEARDSLEY*, Davinia HERNÁNDEZ-LEO & Roberto SÁNCHEZ-REINA
Department of Information and Communication Technologies, Universitat Pompeu Fabra, Spain
*marc.beardsley@upf.edu

Abstract: This paper presents two case studies related to developing educational technologies paired with teacher professional development (PD) courses to support practice-driven research among teachers. The case studies emerged from multidisciplinary projects involving schoolteachers and researchers. The first project supports teacher use of evidence-based teaching strategies that align with the science of learning such as retrieval practice and distributed practice. The second project supports teachers in helping students learn to manage stress using evidence-based self-regulation techniques such as mindfulness and cognitive reappraisal. The projects fuse together concepts of teacher inquiry, learning design, online communities of teachers, and open educational resources to support teacher practice-driven research. An overview of teacher-generated research and impressions of the PD courses gathered during piloting are presented along with descriptions of the two supporting technologies. Implications for practice are also shared.

Keywords: Teacher inquiry, Teacher professional development, Practice-driven research, Learning design, Online communities of teachers

1. Introduction

Practice-driven research is “undertaken in order to gain new knowledge partly by means of practice and the outcomes of that practice” (Candy, 2006). Involving teachers in such research is a way in which “a strong and coherent body of professional knowledge owned by the teaching profession” (Schleicher, 2020, p. 29) can be built. A teacher professional development (PD) approach that aligns with practice-driven research is teacher inquiry (TI) into student learning which involves teachers exploring the use of data to answer questions that emerge from teaching practices (Michos et al., 2018).

Innovative education technologies are well positioned to increase teacher opportunities and abilities to participate in practice-driven research. Learning design (LD) tools, for instance, support teachers in the process of conceptualizing, authoring, and documenting learning tasks (Wasson & Kirschner, 2020). LD tools can be embedded into online communities of teachers (OC) platforms that support interactions among teachers when co-creating, sharing, exploring, and commenting on learning designs in processes of mutual inspiration (Hernández-Leo et al., 2018). The shared learning designs represent open educational resources (OER) that can be freely reused and refined by the educational community. These technologies facilitate collective knowledge building and TI, thereby allowing teachers to expand their practices (Michos & Hernández-Leo, 2020). While there are a number of existing LD technologies focusing on different pedagogical approaches and features, most do not incorporate the required features to support practice-driven research advancing evidence-based teaching opportunities. A system that does integrate options for the required features is the Integrated Learning Design Environment (ILDE). ILDE has been used with online courses supporting PD to improve teachers’ skills when designing for technology-enhanced learning (Garreta-Domingo et al., 2017). With a heightened focus on facilitating TI processes to advance practice-driven research into evidence-based practices among teachers, two projects were undertaken. These projects explore the integration of
innovative educational approaches and technologies such as TI, LD, OC, and OER. The following sections present case studies of the two projects.

2. Case 1: Evidence-Based Teaching

2.1 Problem Definition

“Teachers often use instructional practices known to be wrong (i.e., massing rather than interleaving examples to explain a topic)” (Roediger & Pyc, 2012, p. 242). With increased availability of new technologies (OC, OER) to document and share practices, there is a risk of disseminating ineffective practices. Hence, a multidisciplinary project (Illumine) was initiated to support teacher use of evidence-based teaching strategies and digital technologies.

2.2 Professional Development Course Design

A survey study conducted at the onset of the project, found that teacher motivation is higher for learning to use evidence-based teaching strategies and digital technologies for teaching and learning than for TI for PD (Beardsley & Albo, 2022). To align with these findings, the PD course was designed to focus on introducing teachers to evidence-based teaching strategies and digital technologies to support their use of these strategies. LD and TI were presented as the processes through which teachers would make use of and reflect upon their use of the strategies. A 30-hour course (eleven 2-hour workshops and 8 hours of asynchronous work) was prepared. The first 2 workshops introduced science of learning and learning design theory. Workshops 3-9 alternated between introducing theory and research studies related to teaching strategies and providing time for teachers to co-design implementations. The final workshops involved sharing results and reflecting on the overall course.

2.3 Learning Design Technology to Scaffold Practice-Driven Research

With a focus on introducing strategies and scaffolding TI, the technology (https://illuminne.upf.edu/) is conceptualized as an OC for teachers to learn about and practice their use of evidence-based teaching strategies via the creation of teacher-generated research lessons. A research lesson in this context can be viewed as a simplified interpretation of a Japanese lesson study research lesson (Takahashi & McDougal, 2016) in that teachers conduct the lesson with the purpose of learning about their practices. Research lesson creation is scaffolded via a 5-step teacher-led inquiry design template. TI dictates the selection and sequence of steps with LD providing the functionality for adding learning activities, sharing activity resources, and being able to copy, adapt, and reuse not only the learning materials but the overall research lesson. The reusable research lessons and learning activity material within them are OER contributed to the educational community as shown in Figure 1.

![Figure 1. (L) Illumine research lessons, (R) Remixed open educational resources](image-url)
All in all, the technology scaffolds practice-driven research by teachers on the use of evidence-based teaching strategies through its teacher-led inquiry design template that guides teachers through creating, implementing, reflecting upon, and sharing their research lessons as OER.

3. Case 2: Teaching Students About Stress Management

3.1 Problem Definition

A lack of stress management skills combined with ongoing stressors related to academic demands can negatively impact student academic performance, mental health, and wellbeing (Pascoe et al., 2020). Hence, a multidisciplinary project (RemixED) was initiated to support teacher use of lesson material (OER) introducing evidence-based stress management techniques such as mindfulness and cognitive reappraisal.

3.2 Professional Development Course Design

Results of a survey study conducted at the onset of the project indicate that teachers have a strong interest in learning about the topic of stress (Albó & Beardsley, 2023) but their stress mindsets, defined as their perceptions of the effects and role of stress (Crum et al., 2013), are overly negative and present an obstacle to overcome. Adapting to these insights, the PD course was designed so that participants completed demonstration lessons on stress mindsets and self-regulation techniques as a way to ensure participants learned about positive stress mindsets. These lessons provide an initial experience participants could build upon when adapting additional OER for their students. LD and TI were introduced as processes through which teachers were to plan and reflect upon their use of OER. A 20-hour course (six 2-hour workshops and 8 hours of asynchronous work), shown in Table 1, was prepared.

Table 1. Teacher PD Course on Evidence-Based Self-regulation Techniques (Case 2)

<table>
<thead>
<tr>
<th>Session</th>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>Getting started: Expectations &amp; demonstration lessons</td>
</tr>
<tr>
<td>S2</td>
<td>2</td>
<td>Preparation: Introduction to learning design concepts</td>
</tr>
<tr>
<td>S3</td>
<td>4</td>
<td>Planning: Research lesson sharing &amp; feedback</td>
</tr>
<tr>
<td></td>
<td>6-8</td>
<td>Autonomous work: Conducting research lessons with students</td>
</tr>
<tr>
<td>S4</td>
<td>9</td>
<td>Sense making: Research lesson reflections</td>
</tr>
<tr>
<td>S5</td>
<td>11</td>
<td>Publishing: Presenting work &amp; OER publication</td>
</tr>
<tr>
<td>S6</td>
<td>12</td>
<td>Lessons learned: Reflecting on the overall course</td>
</tr>
</tbody>
</table>

3.3 Learning Design Technology to Scaffold Practice-Driven Research

With a focus on guiding teachers to help their students learn to manage stress using evidence-based self-regulation techniques, the technology (https://remixed.upf.edu/) is conceptualized as an OC for teachers to discover, adapt, and share OER related to stress management (see Figure 1). OC is emphasized through the community interface that facilitates interactions among community members. LD is at the center of the technology in both the creation, sharing, and reuse of learning activities and lessons. Through the LD interfaces, teachers are able to upload files to share, document their adaptations, and provide example use cases. TI reflection prompts work within the activity and lesson creation interfaces to scaffold teacher testing and reporting on the usage of the resources shared with the community. All in all, the technology supports practice-driven research by teachers on the reuse of OER related to stress management for students through the integration of TI reflection prompts into the activity and lesson creation interfaces. These prompts benefit both the creators of resources through aiding their planning and reflecting on their teaching practices, and the community by documenting and sharing examples from the authentic use of the resources.
4. Practice-Driven Research Pilots and Implications

4.1 Research Design

A design-based research approach (Cole et al., 2005) was followed in which the projects’ technologies and PD courses were iteratively developed. In both projects, survey research studies were conducted to inform the initial design of the PD course and technologies (Beardsley & Albó, 2022; Albó & Beardsley, 2023). PD courses were then piloted in multiple European countries (Case 1: Spain, Portugal, Serbia, Estonia; Case 2: Spain, Denmark, Cyprus). In these courses, participants worked in small groups (2 to 4 teachers) to design and implement research lessons as a form of collaborative TI. From the pilots, additional data was collected from course participants through pre- and post-course surveys, reflection activities, focus groups, and artifacts produced in the course (i.e., research lessons). A university research ethics committee approved the research in both projects.

4.2 Data Collection and Analysis

This paper presents data related to the artifacts (research lessons), reflection activities, and focus groups from pilots run in Spain. Secondary and vocational education teachers participated in the courses (Case 1: 10 teachers, Case 2: 13 teachers). In both cases, teachers documented their research lessons using a TI template. The data from these templates is summarized in the findings. Additionally, both cases had teachers reflect on the implementation of the course. In Case 1, an asynchronous reflection activity was run at the midpoint of the course and a synchronous reflection activity was run in the final workshop. For Case 2, an online focus group using a conversational method (Devillard et al., 2012) was held with six teachers at the midpoint of the course and a synchronous reflection activity was conducted in the final workshop. Themes were identified from the reflection activity artifacts, transcripts from the focus groups, and observational notes from the final workshops using a cutting and sorting approach (Ryan & Bernard, 2003).

4.3 Findings: Teacher-Generated Research Lessons

Table 3 provides a summary of research lessons run. Case 1 had four research lessons. Three were between group comparisons in which a traditional lesson (i.e., how the teachers would regularly teach such a class) was compared with a lesson using an evidence-based strategy. The fourth research lesson adopted a within group study approach using pre-post measures.

<table>
<thead>
<tr>
<th>Number of research lessons</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategies explored</td>
<td>Spaced learning, brain breaks, reframing, retrieval practice, and distributed practice</td>
<td>Reframing, square breathing, progressive muscular relaxation, mindfulness</td>
</tr>
<tr>
<td>Data collection</td>
<td>Online tools (Google Classroom, Google Forms, Kahoot!), paper-based assignments, and teacher observations</td>
<td>Online tools (Padlet, EdPuzzle, ClassMood App, Google Forms, Google Jamboards, Google Classroom, Kahoot!), paper-based assignments, and teacher observations</td>
</tr>
<tr>
<td>Students participating</td>
<td>~165</td>
<td>~200</td>
</tr>
</tbody>
</table>

From Case 2, six research lessons were run. All followed a within group study approach. All these lessons introduced theory on the science of stress as well as self-regulation techniques. Videos and presentation slides were the most frequently used OER. Most lessons used pre-
4.4 Findings: Teacher Reflections

Challenges mentioned by participants that future iterations of the PD courses need to address are summarized in Table 4. In reflecting on the courses centered on research lessons, Case 1 participants shared how they found the strategies suitable and effective to use with their students based on measures of student performance and/or changes in classroom climate. In writing about what they would change in their future research lessons, participants would adopt a more systematic and rigorous application of the activities over time, include student feedback on their impressions of the activities, and change how they implemented specific strategies such as giving immediate feedback on retrieval practice activities. Overall, the use of strategies in research lessons generated insightful reflections among participants seemingly allowing them to contrast their experience with the training and the provided materials.

Case 2 participants noted benefits to the research lessons such as higher motivation, better performance, and an improved classroom climate. They highlighted changes in the perceptions and attitudes towards stress among students. When reflecting about improvements and future implementations, participants highlighted the importance of expounding the scientific evidence behind the practices; and having colleagues share their positive experiences gained from using the resources. In sum, having teachers perform research activities to not only plan their lessons but also reflect on lesson outcomes in a collaborative manner was positively received.

<table>
<thead>
<tr>
<th>Themes</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload and time</td>
<td>Having to learn theory, create and implement research lessons was too</td>
<td>A lot of time needed to be invested to design and then implement sessions.</td>
</tr>
<tr>
<td>pressure</td>
<td>much work. Also, time to collaborate was not made available within their</td>
<td>Also, the course was cognitively demanding as a lot of information</td>
</tr>
<tr>
<td></td>
<td>current teaching schedules</td>
<td></td>
</tr>
<tr>
<td>School culture</td>
<td>Peer work was not part of the school culture.</td>
<td>A culture of sharing in which time is available to analyze and discuss</td>
</tr>
<tr>
<td></td>
<td></td>
<td>materials and experiences was not yet established in the school.</td>
</tr>
<tr>
<td>Bridging theory and</td>
<td>It was a challenge to translate research studies on teaching strategies</td>
<td>NA</td>
</tr>
<tr>
<td>practice</td>
<td>into practice</td>
<td></td>
</tr>
</tbody>
</table>

4.5 Implications for Practice

In the context of practice-driven research, PD course participants were able to learn about evidence-based teaching, come up with inquiry questions related to the introduced strategies and materials, design learning activities to explore their questions, collect data through implementing the activities with students, reflect on their experiences, and produce lessons learned. In a sense, they were able to conduct practice-driven research and generate outputs that can be useful to the educational community – as case studies of their experiences. Nonetheless, there are still gaps to close before such work can make greater contributions to professional knowledge owned by the teaching profession. Similar to obstacles identified that have slowed the adoption of Japanese Lesson Study outside of Japan (Chokshi & Fernandez, 2004), we observed that many participants struggled to free up enough time to adequately engage in inquiry processes and collaborate with peers, lacked procedural research knowledge and data literacy that affected the quality of data collected and their interpretations of the data, and needed to use the resources (i.e., strategies and materials) with students more than one time to become proficient in using them comfortably for research lessons.
In terms of implications for practice and policy, interpretations of these two cases suggest that collaborative TI, evidence-based teaching, and digital technologies are themes that can be used to engage teachers in continuing PD. Yet, to have teachers contribute to a professional knowledge base grounded in evidence through practice-driven research, efforts are needed to overcome gaps in pedagogical and inquiry practices (e.g., procedural research knowledge and data literacy) and to lessen perceived burdens (workload, time pressure, school culture). Ongoing PD that supports teacher development of both pedagogical and inquiry practices in parallel is likely to play a critical role in establishing and maintaining teacher-owned professional knowledge. Further, innovative technologies that foster online communities of teachers, scaffold LD and TI processes, provide reusable research lessons (OER) – such as the examples presented in this paper – have the potential to support and sustain teacher PD that targets both practices.

Acknowledgements

The authors would like to thank Illumine and RemixED project members and participants for their contributions. This work has been partially funded by Erasmus+ (2020-1-ES01-KA201-082504, 2021-1-ES01-KA220-SCH-000032801), Department of Research and Universities of the Government of Catalonia (SGR 00930), and MICIN /AEI/ 10.13039/501100011033 (PID2020-112584RB-C33).

References


The Impact of Preschool Teacher Training on STEAM Education: Professional Preparedness and Confidence

Wan-Chen CHANG
Department of Early Childhood Education, National Tsing Hua University, Taiwan
*wc_chang@mx.nthu.edu.tw

Abstract: This study aims to explore the impact of teacher training in the field of STEAM education on preschool teachers' preparedness and confidence. The participants of this study were nine education and childcare professionals from a public preschool in northern Taiwan. A small-scale workshop was conducted to train the teachers, covering topics such as an overview of STEAM education, STEAM learning for preschoolers, teaching methods, and curriculum design. The results revealed that prior to the training, the teachers had a high awareness of the importance and expected outcomes of STEAM education. However, they scored relatively low in terms of teaching experience and self-efficacy. After the training, the teachers showed significant improvement in their knowledge of STEAM teaching methods and environmental support. Nevertheless, their ratings for confidence and teaching efficacy remained relatively low. These findings highlight the need for further professional development and support for teachers in the field of STEAM education. The results of this study provide valuable insights and recommendations for teacher training and instructional practices in preschool settings.

Keywords: preschool teacher training, STEAM education, confidence, teaching efficacy, curriculum design

1. Introduction

Under the impetus of policies and educational trends, both parents and teachers are highly supportive of implementing STEAM education at the preschool stage. However, adults need additional knowledge and support to facilitate the execution of STEAM education (e.g., Margot & Kettler, 2019; McClure et al., 2017; Simoncini & Lasen, 2018). In particular, teachers play a crucial role in STEAM education at the preschool level (Karademir et al., 2020; Simoncini & Lasen, 2018; Monkeviciene et al., 2020; Wahyuningsih et al., 2020). This is because teachers can only successfully incorporate it into the curriculum when they fully understand the components of STEAM education (Pang & Good, 2000). However, to successfully integrate STEAM education into the curriculum, teachers need to grasp additional knowledge and skills. Therefore, it is vital to provide training and support for preschool teachers' professional development in STEAM education. This study aims to understand preschool teachers' cognition and beliefs about STEAM education, their professional needs, and challenges in STEAM education, and to explore how to provide effective professional development training and support in STEAM education. The goal is to enhance preschool teachers' professional capabilities and confidence in STEAM education, thereby promoting children's learning and development.

2. Research Method

2.1 Research participants
2.2 Research Instruments

2.2.1 STEAM Teaching Questionnaire for Teachers
This study will develop a questionnaire on teachers’ beliefs about STEAM teaching based on previous research. Pre-tests will be administered when the teachers participate in the first phase, and post-tests will be administered after the second phase concludes. The questionnaire structure includes five major dimensions: the importance of STEAM education, teaching methods for implementing STEAM education, support for implementing STEAM education, self-efficacy, and expected outcomes. The questionnaire uses a seven-point scale, with respondents selecting an option from a continuum ranging from "strongly agree" to "strongly disagree" (7 6 5 4 3 2 1).

2.3.2 Mid-term Progress Review of Teachers’ STEAM Teaching
To understand how teachers are implementing the STEAM curriculum during the semester, the researchers will have discussions and interviews with the teachers from each class during the semester to understand the challenges they face while implementing the curriculum.

3. Preliminary Data Analysis

This study analyzed the data from the teachers before the training, and the results are shown as below. Looking at the means, teachers have a positive attitude in all five dimensions. Particularly, teachers believe in the importance of STEAM teaching (M = 5.65, SD = 1.09) and have a higher average score for the expected outcomes of STEAM teaching (M = 5.58, SD = 1.06), indicating that teachers have a more positive attitude towards the learning outcomes of STEAM teaching for children. The average score for teaching experience is relatively low (M = 4.23, SD = 1.31), indicating that teachers' previous experience with STEAM teaching is quite lacking, even including a lack of experience in science inquiry teaching.

Further analysis on the importance of STEAM activities indicates that teachers believe effective STEAM education is crucial for enhancing children's hands-on capabilities and problem-solving skills (M = 6.19, SD = 0.71), as well as stimulating their curiosity and interest in everyday phenomena (M = 6.10, SD = 0.79). On the other hand, they do not place as much emphasis on enhancing children's acquisition and understanding of STEAM knowledge (M = 4.67, SD = 0.92).

In response to open-ended questions, the majority of teachers (85%) believe that the essence of STEAM education lies in using inquiry-based problems to help children develop relevant skills, such as learning by doing, logical thinking, problem identification, problem-solving, and collaborative abilities. Regarding potential future challenges, most teachers (94%) are concerned about their lack of STEAM teaching expertise, particularly in preparing teaching methods. This result aligns with the findings from the questionnaire.

After the implementation of the thematic courses, a mid-term review revealed that teachers still lack confidence in their teaching (M = 3.94, SD = 1.27). The teachers’ efficacy in STEAM teaching (M = 4.29 , SD = 1.20) and encouragement for children to use scientific inquiry methods in STEAM teaching (M = 5.04, SD = 1.35) were also examined. This shows that, compared to before the training, teachers are now able to utilize more scientific inquiry methods, but there has been no significant difference in their teaching efficacy. Therefore, it is necessary for university professors to provide more explicit guidance to give them concrete goals for STEAM teaching.
4. Conclusion and Discussion

The results of this study indicate that teachers have high expectations for the implementation of STEAM education in early education stages, and they highly value the development of children's abilities. After the training courses, teachers utilized more child-centered teaching methods in STEAM courses, but their self-efficacy in teaching remained insufficient. In addition, teachers moderately utilize exploratory methods and dialogue to foster the cultivation of higher-order thinking skills, as well as reflection and self-reflection abilities (Monkeviciene et al., 2020). Therefore, professional development in STEAM domain knowledge and teaching methods are necessary for teachers in STEAM education.

Acknowledgements

The authors would like to thank National Science and Technology Council of the Republic of China, Taiwan, for financial support (NSTC 111-2410-H-007-063 - )

References


Matching Intervention Messages Considering Complex Personality Types of High School Students

Taisei YAMAUCHIa, Yuta NAKAMIZOb, Kyosuke TAKAMIC, c, Rwitajit MAJUMDARc, & Hiroaki OGATAb, c

aGraduate School of Informatics, Kyoto University, Japan
bEducation Data Science Center, National Institute for Educational Policy Research (NIER), Japan
cAcademic Center for Computing and Media Studies, Kyoto University, Japan
* yamauchi.taisei.28w@st.kyoto-u.ac.jp

Abstract: In this paper, we proposed intervention messages that consider the complexity of student personalities. Because of the increasing importance of individually optimized learning, the intervention for the student is also desired to be tailored. While previous studies have found the appropriate message content related to one or several personality traits, no study has been conducted to send individually optimized messages with consideration of each student’s complex personality traits. In this study, we conducted the clustering of students based on Big Five test scores and then determined the nature of each cluster and the type of intervention message related to that nature. According to the clustering analysis, the students were categorized into three groups. Considering the personality trait of these groups, we decided tailored messages for each group.

Keywords: Message intervention, big five inventory, individually optimized learning, personality, classification

1. Introduction

The use of the flaming intervention is the introduction of nudges in educational studies (Damgaard & Nielsen, 2018), and it can be an effective way of individually optimized message intervention. Some studies using nudge interventions insist on the improvement of students’ performance (Motz et al., 2020; O’Connell & Lang, 2018). While previous studies show that flaming intervention can be applied to individually optimized message intervention (Wall et al., 2019; Yamauchi et al., 2022) using Big Five Inventory (John et al., 1991), these studies focus on only one or several types of personality traits, and they did not propose the specific message type for each cluster.

Hence, the aim of our research is to suggest an approach for identifying effective intervention messages considering students’ complex personality traits. We try to investigate the following research question: what are the clusters of complex personality types of high school students and what is the optimized message for each of those clusters?

2. Literature Review

2.1 Big Five Inventory

Personality inventories are a common approach for evaluating personalities in psychology. They involve a series of questionnaires designed to uncover a subject’s personality. One well-known inventory is the Big Five Inventory (John et al., 1991), which categorizes personalities into five groups. This method is highly popular, and researchers have even developed lists of appropriate adjectives for each category (Goldberg, 1992; Hofstee &
Raad, 1992; Johnson & Ostendorf, 1993). The relationships between each Big Five trait and adjectives are represented in the following items, where (+) means the profile those who have the high Big Five score are likely to have, and (-) means the ones those who have the low Big Five score are likely to have:
- Openness to experience: inventive (+), curious (+), consistent (-), cautious (-)
- Extraversion: active (+), talkative (+), shy (-), quiet (-)
- Agreeableness: kind (+), cooperative (+), selfish (-), cold (-)
- Conscientiousness: responsible (+), organized (+), extravagant (-), careless (-)
- Neuroticism: anxious (+), angry (+), calm (-), unemotional (-)

In our research, we will use the Japanese Big Five Inventory to determine each student’s personality.

2.2 Message Interventions in Educational Context

In education studies, the use of nudges has received much attention (Damgaard & Nielsen, 2018). Nudge is defined as “alter[ing] people’s behaviors in a predictable way without forbidding any options or significantly changing their economic incentives”, which is the term of economics (Thaler & Sunstein, 2008). A study suggested that students do not mind receiving nudges more often if they perceive the nudges to be useful to them (Gatare et al., 2021). The introduction of nudges in education studies is called a framing intervention. Even small changes in the framing of information have the potential to alter behavior and eliminate biases due to cognitive and attentional limitations (Damgaard & Nielsen, 2018). In this study, we try to use flaming intervention for an individually optimized system.

It is possible to create various types of nudges that consider the student’s personality, which may have varying impacts on how students approach their quizzes. One study determined on an individual basis which messages are most appropriate based on profile characteristics (Yamauchi et al., 2022). According to this study, there existed a noteworthy inverse association between the rate of responding to notifications and conscientiousness for peer notifications. Additionally, there was little favorable correlation between the frequency of notification responses and conscientiousness for deadline notifications. Moreover, neuroticism was correlated with peer notifications, while openness to experience was correlated with peer notifications as well.

Although some researches show the effectiveness of intervention related to one type of or several types of personality (Wall et al., 2019; Yamauchi et al., 2022), no study shows the application of individually optimized content message considering each student’s all complex personality. It is difficult to decide what type of message should be sent to each student because every student has a complex personality. In this study, we challenged the clustering of the students in terms of their personality and made a decision of sending what type of message and then apply it to students.

3. Research Design

This study was conducted in a high school in Japan. 279 1st-grade students (15-16 years old) participated in this experiment. To estimate students’ personalities, we used a Japanese version of a questionnaire survey which consists of 60 questions about their personal traits (Murakami & Murakami, 1997). This can be helpful when someone takes the survey to those who cannot understand the original English questionnaire. The Big Five assessment comprises 70 questions, out of which 60 are presented as true/false questions relating to one of the five personality traits. Each student’s personality score is determined by adding up the scores for each item on a 12-point scale by the student’s answers. We collected Big Five personality data from 120 students of all 279 in this experiment.

We designed and developed a message-sending system called “LRS2Message”, which combines Learning Record Store (LRS), Learning Management System (LMS), and Analysis Tool. Figure 1 describes the chart of to whom the system sends a message and how to send messages to the environment. We schedule the message token to store the message
logs into LRS by pushing the checkboxes and a specific button in Analysis Tool (surrounded by a red square in Figure 1). On the other hand, we used LRS2Message to set some emails with sending contents, timings, and dates. Once the time has come, LRS2Message gives LRS the query to decide to whom the system sends a scheduled message. LRS2Message also gets the result of the query, and then it sends a message to LMS according to the result of the query. We can also send a message manually with Analysis Tool, so we used it to send a part of the message.

![Message-sending flow using LRS2Message](image)

Because of the many variables of personality traits, we can’t easily decide with some comprehensive rules what type of contents of messages the student receives. Therefore, we conducted the clustering with Big Five test scores of five personality traits as explanatory variables and then we decided the proper message contents of each cluster instead of making rules that can assign property to each student. Based on the result of the k-means elbow plot of that clustering, we found that 3 clusters seem the best number of clusters because it is the point at which the absolute value of the change in slope is greatest. Previous studies follow this result that we can classify the students into 3 clusters (Asendorf & van Aken, 2001, 1999; Robins et al., 1996; Takami et al., 2023; Wall et al., 2019).

4. Findings

Based on the result of the elbow plot, we clustered 120 students into three clusters. Table 1 shows the cluster name, the number of people in the cluster, the score means and SD of students’ personality traits of each cluster, and the graphs of the distribution of students’ personality traits. According to these results, we named these clusters whose names represent the property of each cluster. The first cluster has a higher-than-average score for all five personality traits, so we named this cluster “Hi5”. The second cluster has a higher-than-average score for agreeableness and extraversion and a lower-than-average score for neuroticism. We named this cluster “AEn” based on the characteristic score. The third cluster has a higher-than-average neuroticism score and has lower than average openness to experience or extraversion score. According to the features, we finally named the cluster “Noe”. We also found from ANOVA analysis of the data that all of the Big Five personality scores of one specific cluster are significantly different from the other ones. This result fastens the validity of the assignment of the properties of each cluster.

At the second step, we assigned which message to send to each group of the cluster. Students in the Hi5 group have the highest average C score of the three clusters. One previous research shows that it is better for students’ engagement to send deadline-type message to people who have a high C score (Yamauchi et al., 2022). Thus, we think it is best to send a deadline-type messages to them. Students in the Noe group have the lowest average C score of the three clusters. One study also shows that it is better for students’ engagement to send peer-type messages to people who have a low C score (Yamauchi et al., 2022). Thus, we concluded that it is best to send peer-type messages to them. Students in the AEn group have the lowest N score of all three groups. Students who have a low N score are unemotional
(Hofstee and Raad, 1992) and calm (Johnson & Ostendorf, 1993). Thus, we decided to send messages which have the content on prior actions of herself/himself.

Considering the personality trait of each message type, we decided what contents of the message are the best to send in this engagement context. The following sentences are examples of the content:

- **Deadline:** “You have not worked on the recommendation system or test set maker in the last 3 days. The exam will begin on February 24. The exam date is only 11 days away. AI recommendation will help you learn more efficiently.”

- **Peer:** “You have not worked on the recommendation system or test set maker in the last 3 days. During the period, the recommendation system or test set maker was used by 7 people, and 15 exercises were solved. AI recommendation will help you learn more efficiently.”

- **Commitment:** “You have not worked on the recommendation system or test set maker in the last 3 days. You have solved one recommendation exercise so far. You need to solve 4 more exercises to use Test Set Maker and further improve your academic performance. AI recommendation will help you learn more efficiently.”

Table 1. Classification property

<table>
<thead>
<tr>
<th>Type (# of people)</th>
<th>Mean (SD) Score of each personality trait</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Hi5 (48)</td>
<td>5.98</td>
</tr>
<tr>
<td>Noe (31)</td>
<td>3.27</td>
</tr>
<tr>
<td>AEn (41)</td>
<td>4.68</td>
</tr>
</tbody>
</table>

5. Discussion and Limitations

In order to provide appropriate message interventions that take into account complex personalities, clustering was performed on the students and the characteristics of each cluster were looked at to determine appropriate intervention messages. The results showed that the best approach was to divide the students into three clusters, and we also found that the appropriate intervention message type for each is “Deadline,” “Peer,” or “Commitment” based on previous literature (Wall et al., 2019; Yamauchi et al., 2022). Following notes are the features of each cluster:

Group “Hi5”: This group has a higher C score than the others, and previous study shows people with higher C score tends to be punctual (Back et al., 2006). One study has shown that the more deadlines there are, the better students tend to perform (Rabin & O’Donoghue, 1999). From that point of view, Hi5 students can be susceptible to a deadline-type message.

Group “Noe”: The other group has a higher N score and lower E score than that of the others, and previous study shows that people with this personality are obeying those in authority, and they fear social disapproval (Wall et al., 2019). Previous studies (Carrell et al., 2013; Papay et al., 2020; Rogers & Feller, 2016) have shown that intervention which encourages peer-to-peer collaboration can be effective in promoting a sense of social belonging, establishing social norms for striving, and improving skills through partnerships.

Group “AEEn”: This group has a higher A score and a lower N score than that of the others, and previous studies showed that people with a higher A score and a lower N score are likely to maintain consistent beliefs they are owning (Cialdini, 2001; Wall et al., 2019). They are more susceptible to commitment persuasion strategy. Those who have strong social skills are more likely to be influenced to perform a behavior if it aligns with their existing beliefs or prior actions (Cialdini, 2001; Wall et al., 2019). Thus, we selected to send a commitment-type message to them.
One limitation of the experimental result is that it is not necessarily true that everyone can adapt the cluster’s rule. For example, we can see from Table 1 that there is a student who has a 1-point C score even in H15. For the solution of these problems, we could use more indicators to evaluate students’ personality like Dark Triad Traits (Jonason & Webster, 2010) or Type-D (Denollet, 2005).

6. Conclusion and Future Works

In this paper, we suggested intervention messages considering each student’s complex personality and presented techniques for visualizing the outcomes using relevant indicators extracted. First, student personalities were investigated using the Big Five test, and then students were clustered using it. The clustering results and previous papers were then consulted to determine the nature of each cluster and the type of intervention message adapted to that nature. The clustering results showed that the students were divided into three groups, and we concluded that the best messages for each of these groups were “Deadline,” “Peer,” and “Commitment”, respectively. Previous works of literature follow the results of this clustering and the validity of the appropriate intervention messages.

We also conducted an intervention experiment with the same cohort of students to understand the effects of matched and random messaging. However, due to the small number of student engagement logs, the result is not conclusive and would require further study.

Acknowledgements

This research was supported by JSPS JP22H03902, JP23K17012, NEDO Special Innovation Program on AI and Big Data JPNP18013, NEDO JPNP20006.

References


Design-Based Implementation Research: A Collaborative Approach to Educate Out-of-School Children

Faisal BADAR* & Jon MASON
Charles Darwin University, Australia
*Faisal.Badar@students.cdu.edu.au
Jon.Mason@cdu.edu.au

Abstract: Design-Based Implementation Research (DBIR) is characterised by iterations of a solution informed by collaboration with stakeholders and often used in contexts involving digital technology. It is also useful in building the capacity for long-term collaborations between research and practice for sustainable and equitable change within an educational system. In this study, DBIR has informed a scalable solution to provide quality and inclusive education to underprivileged out-of-school children, using the Digitally Operated One-Room Schoolhouse (DOORS) model. In doing so, it responds to questions of importance in the local context and builds the capacity of systems to engage in continuous improvement. This paper presents a summary.

Keywords: Design-Based Implementation Research, out-of-school children, education, design thinking

1. Introduction

Design-Based Implementation Research (DBIR) is a research approach based on principles of Design-Based Research (DBR) that foreground the role of digital technologies in improving educational practices through an iterative, collaborative process with key stakeholders (Amiel, & Reeves, 2008; Anderson & Shattuck, 2012). SDG-4 is one of the UN’s Sustainable Development Goals that aims to provide universal primary and secondary education. However, as of UNESCO’s 2022 report, there are still an estimated 244 million children aged between 6 and 18 years out of school globally. Over the years, global stakeholders have launched several initiatives such as Education for All (EFA), the Millennium Development Goals (MDG) and One Laptop per Child (OLPC) to ensure the provision of primary education to all children. However, these initiatives were unsustainable in meeting their objectives and their lack of delivery of expected progress may be directly attributed to the failure of considering contextual factors in shaping such initiatives (Badar & Mason, 2020). As an iterative and collaborative approach to developing solutions, DBIR can assist in adaptable, scalable, and sustainable outcomes for educating OOSC.

This paper reports on a DBIR-based educational initiative that utilises the integration of education technology tools and the one-room schoolhouse concept. The Digitally Operated One-Room Schoolhouse (DOORS) model implements an innovative and overarching curriculum and pedagogy to provide inclusive and quality education to OOSC using blended learning techniques. This program progressed pragmatically from exploratory study to design-oriented research, in which feasibility and fine-tuning of learning environments within rural and remote location contexts was the focus. It then progressed towards rigorous collaborative efforts among all stakeholders including researchers, practitioners, OOSC parents, influencers around OOSC, and policymakers. The iterations were intended to respond to deeply entrenched problems and pragmatically implement an adaptable, sustainable, and scalable system to educate underprivileged OOSC, suited to their local challenges and limitations.
2. EDvantage Digital Learning System-A DBIR Approach to Educate OOSC

EDvantage Digital Learning System (EDLS) is a Digitally Operated One-Room Schoolhouse (DOORS) model, which is a result of extensive research and an iterative implementation process based on the DBIR approach. EDLS is a complete schooling system, built through the collaboration of researchers and practitioners over the period of the last five years. The project is currently operational in remote and rural areas of Pakistan and Tanzania, providing free, inclusive, and quality education to OOSC through the integration of modern digital educational technologies with the centuries-old one-room schoolhouse concept, suited to their local challenges and limitations. The EDLS-DOORS’s basic design and implementation process follows design thinking basics as shown in Figure 1 below.

The above six-phase process was not implemented linearly but rather iteratively, especially during the first six months of the prototype phase of the pilot campus. After completion of six months iterative process of the pilot campus. The subsequent development and growth from a single prototype campus to 14 campuses during the next 20 months. The EDLS-DOORS model adopted a pragmatic approach and conduct a series of iterations to implement best-suited academic and pedagogical processes as shown in Figure 2.

![Figure 1: Design and Implementation Process of EDLS-DOORS Model.](image)

**Figure 2: EDLS-DOORS Iterative Model using DBIR Approach**
The EDLS_DOORS Model consists of three series of iterative stages under a DBIR framework – the pilot campus phase for concept testing, the multiple campuses phase for adaptability and sustainability. The first series of iterations involved the initial testing of the concept through the establishment of the pilot campus in 2021. From the initial idea generation in 2018 to the launch of the pilot campus in 2021, initial planning and research were conducted to understand the OOSC problem, gain familiarity with contextual factors of underprivileged OOSC through community visits, gauged the local sentiments towards education for feasibility analysis, studied academic best practices suitable for the contextual requirements, and understand the EdTech tools suitable for underprivileged OOSC education. The pilot campus, which was crucial for testing the EDLS-DOORS concept's applicability and identification, was launched in a rural area of Karachi, Pakistan and observed closely for six months and iterations were continuous and extensively applied to improve the overall model.

The second series of iterations involves adaptability and sustainability testing through multiple campuses at different locations. The understanding of unique local contextual factors of OOSC in each location was established for ten new campuses in rural areas. Through site visits and meetings with community stakeholders, the local dynamics and limitations are understood and required adjustments to the model were made. Issues such as language barriers, parental reluctance and cultural restraints were understood and applicable solutions were designed, making the model sustainable and adaptable.

The third series of iterations involved testing the scalability of this model through expansion to multiple regions outside of Pakistan to understand and focus on local contextual factors of OOSC in different regions of the world. Reaching out to and collaborating with local stakeholders in those regions was found to be crucial and yet stakeholders from five different countries approached. The scalability testing in this iteration has been initiated through the launch of online Spoken English classes in rural Tanzania.

3. Conclusion

DBIR is a systematic approach for working with stakeholders to iteratively design research that focuses on implementation that responds to questions of importance in the local context. As outlined in this paper DBIR has been pivotal in guiding the early success of the EDLS-DOORS implementations. DBIR poses its own set of challenges and limitations to the researcher. Establishing researcher-practitioner partnerships and building trust may bring along a challenge for researchers, particularly with new stakeholders with whom no previous relationship exists (Fishman, Cheng, & Penuel, 2019). EDLS-DOORS model provides an opportunity to approach out-of-school children’s education differently and shows promising outcomes in terms of its adaptability, sustainability, and scalability.

References

A Study of Versatile Tutor Training Programs for Universities in Japan

Yumi ISHIGE*, Kazuhiro KABEYAb, Kayoko NAGAOc & Hirotoshi TANIGAWAd

*Modern Social Studies, Otemae University, Japan
b School of Health and Sport Sciences, Osaka University of Health and Sport Sciences, Japan
c School of Health and Sport Sciences, Osaka University of Health and Sport Sciences, Japan
d Department of Early Childhood Care and Education, Shikoku University, Japan
*y-ishige@otemae.ac.jp

Abstract: This research reports on the creation of online tutor training programs for learning support centers in Japanese universities. As this topic is relatively new in Japanese higher education institutions, there is a lack of practical training programs that can serve as references. To address the issue, we have opted to develop programs based on a framework established in the United States. The programs were planned in accordance with two core concepts: “usable by any university” and “to be self-learning materials.” The primary objective was to create training programs that can be universally used in Japanese universities, allowing tutors to learn independently even without a lecturer. In conclusion, it can be affirmed that the programs serve as versatile tutor training resources. While some cultural differences may have emerged due to the adoption of a US-based framework, it is believed that these variances can be partially resolved by making adjustments during on-site programs.

Keywords: tutor training program, learning support center, online learning materials

1. Introduction

The main propose of this research is to develop training programs for tutors based on a framework used in the United States and to validate the effectiveness of the programs for Japanese universities. There are two primary reasons for undertaking this research. First, the necessity of a learning support center (LSC) in higher education has been recognized in Japan since around the latter half of the 2000s (The Japan Association for Developmental Education, 2012). In this research, a LSC is defined as an organization that assists students with their learning outside of classroom (Ministry of Education, Culture, Sports, Science, and Technology, 2010). In light of this need, the importance of training programs for tutors working at LSC has gained attention recently. However, there are currently no authorized and established tutor training programs in Japan, despite individual institutions implementing their own programs. Several LSCs appear to require more information on tutor training. The creation of such programs is expected to improve the situation better.

Second, there is a significant disparity between the abundance of tutor training materials available in the United States compared to Japan. While these resources can serve as valuable references, it is challenging to adopt them directly without modifications. This is because the requirements for academic support in Japanese universities are perceived to differ from those in the United States. Thus, it becomes necessary to adapt the materials to suit the Japanese context. Therefore, we made the decision to design our own tutor training programs referring to the research of The College Reading and Learning Association (CRLA). CRLA is known for issuing certifications for international tutor training programs, making it suitable for our purposes. This paper will present the contents of the tutor training programs we have developed and examine their effectiveness.
2. Method: Creating tutor training programs

Two key concepts guided the creation of the tutor training programs in this research: one is “usable by any university” and the other is to be “self-learning materials.” While it is undeniable that each university, faculty, and department has its own study styles and learning contents, it is unlikely that a single teaching method would be universally suitable for all subjects. However, we believe tutoring skills aimed at supporting tutees can be developed regardless of the differences between universities, instead of focusing on teaching methods for specific subjects. Tutor training programs that we aim to develop should adhere to these principles. Secondly, it is important to note that not all universities in Japan with LSC have a lecturer who can train tutors. The reasons for this absence may vary depending on the university, such as budget constraints, lecturers being occupied with other responsibilities, and/or a lack of understanding of training needs. In order to make the situation better, providing self-training materials can be an effective solution. Thus, we have decided to develop materials that can be utilized for both collective and individual learning.

Based on the aforementioned principles, we have selected the article “Standards, Outcomes, and Possible Assessments for ITTPC Certification” by CRLA as basic reference. ITTPC stands for “International Tutor Training Program Certification,” which was established to provide “recognition and positive reinforcement for tutoring program’s successful work from an international organization,” and it also requires “set an internationally-accepted standard of skills and training for tutors” (The College Reading and Learning Association, 2018). The article presents ideas on tutoring skills that are applicable internationally, rather than specific to particular fields of study. We anticipate that it will be well-suited for our intended purpose. While our research does not aim to develop programs by which the certification can be acquired, the standards can be a good example what tutors are required at higher education.

Five researchers, including myself, have actively participated in the research. We engaged in extensive discussions and evaluations of the contents of ITTPC and of our programs, then we were assigned to topics and made each of them, and brought them in for further discussion. One of the distinguishing features of the programs is that they were designed as online materials, allowing for easy accessibility and reference. Consequently, we refer to these online tutor training programs as “OTT.” Each topic of OTT consists of a short video, a note containing key sentences, and a worksheet with assignments. These assignments consist of both individual and multi-person tasks. The note with key sentences and the worksheet with assignments are provided as downloadable PDF files, allowing for easy access and use.

3. Result

Tutor training at Otemae University was conducted even before the existence of OTT, during which a diverse range of materials were utilized based on the identified necessary topics. However, compared to that time, the implementation of OTT brings several advantages. First, having a single set of training materials enables both tutors and lecturers to adhere to a consistent policy, providing a sense of stability and security. The establishment of OTT has indeed resulted in increased reliability, which is believed to be associated with the reliability of ITTPC. Tutors have also provided feedback stating that the expectations placed upon them have become clearer through the use of OTT.

Second, tutors use worksheets with assignments to help them in preparing their workshops during the training. Those materials can be valuable resources for tutors, assisting them in organizing specific segments of tutor training. Third, the availability of online materials has made it easy for tutors to work on preparation and review, and create-up work for absent students is easier than when there were no such materials available.

Although the benefits of OTT have been highlighted above, some challenges have emerged. OTT’s universal design occasionally caused confusion among tutors as certain descriptions did not align with their specific LSC contexts. Moreover, misunderstandings of assignment intent led to workshops being perceived as ineffective. These issues currently
pose negative aspects of OTT. Another challenge lies in the widespread dissemination of OTT to Japanese universities.

4. The effectiveness of OTT

To sum up, topics offered by ITTPC serve as an effective framework for tutor training at Japanese universities. The existence of such programs has been helpful, allowing students to develop their skills and awareness through the use of OTT. The well-designed framework serves as a valuable reference for tutors. Even if a lecturer does not use OTT directly for their own tutor training, presenting the topics of OTT as a guide to the required skills and competencies of tutors can still be effective. It is worth noting that OTT is available in Japanese, further enhancing its accessibility and relevance in the Japanese context.

Nonetheless, the impact of cultural differences on the use of OTT cannot be completely ignored. In the case of Otemae University, a topic “Adult learners” was a notable example. In Japan, the majority of university students enroll immediately after graduating from high school, resulting in a limited number of mature students. For instance, at my university, there are only one or fewer mature students per grade level who utilize the LSC. Thus, the topic is not used in tutor training at Otemae University as an opportunity to take advantage of is really rare. This situation is not unique to Otemae University but is common among many universities in Japan (e-Stat, 2021). Another case is “Supplemental Instruction.” While the instructional method is explained in the material, it is not typically organized or utilized in the context of Japanese higher education institutions. Therefore, at Otemae University, Supplemental Instruction is also omitted from tutor training due to its limited applicability in the typical Japanese educational setting.

Hence, it is evident that not all content in OTT is invaluable to a specific university. This limitation arises from the nature of versatile programs. However, it is expected that this issue can be addressed through the way individual implementation and operation of OTT by each university. Overall, it can be concluded that the objective of creating versatile tutor training programs using the ITTPC framework has been achieved, and significant results have been obtained.

In the future, we have plans to improve the content by addressing any aspects that may lead to misunderstandings among tutors. Besides, as OTT is a freely accessible online resource, we aim to collaborate with educational societies to promote its widespread use. While we have only presented the practices of one university in this study, our intention is to gather operational data from multiple universities and conduct future comparisons.

Acknowledgements

This work was supported by Grants-in-Aid for Scientific Research 17K01163 and 20K03112.

References


Support for fitting Chromebooks to the child with cerebral palsy: A practical study on incorporating advice from ICT specialists

Tomohito YAMAZAKII* & Toyokazu MIZUUCHIB+)

a Faculty of Economics, Asahikawa City University, Japan
b) Faculty of Human and Culture, The University of Shimane, Japan
*yamazakitmht@live.asahikawa-u.ac.jp

Abstract: An ICT specialist provided guidance for a fitting process aimed at enhancing Chromebook usability for a child with cerebral palsy. By gathering feedback from the children and observing their interactions with the device, adjustments were made using accessibility functions to improve ease of use. Expert advice addressed challenges beyond accessibility improvements, resulting in a successful fitting process.

Keywords: cerebral palsy (CP), Chromebook, fitting, accessibility, special support education

1. Introduction

The GIGA School Initiative has provided one device per student in schools across Japan, including those for special needs. A report indicates that ChromeOS constitutes 40.1% of these devices, implying widespread use. However, adapting Chromebooks for students with cerebral palsy is challenging due to their unique physical limitations. Tanioka et al. (2008) stated, "In general, physically disabled people with cerebral palsy may have involuntary actions, so when they consciously try to move one point on their body, they may not be able to move it properly due to the force of their entire body." He mentioned that a system that enables text input and PC operation by voice has been developed. However, there are differences in the devices and their OS installed in special-needs schools in each of the local municipalities, and the state of children itself varies, it is difficult to construct a system tailored to each child as Tanioka has suggested. In addition, lack of teachers' knowledge of ICT devices may lead to the possibility that many special-needs schools do not take advantage of the accessibility functions of their devices. In this study, therefore, with the advice of experts, a fitting was conducted using accessibility functions to make the Chromebooks easier to use for children with cerebral palsy.

2. Purpose & Method

2.1 Purpose

The purpose of this study aimed to enhance Chromebook usability for a child with cerebral palsy through expert-guided customization, focusing on the fitting process to facilitate the child's effective usage.

2.2 Subject

The study focuses on X, a junior high school student with cerebral palsy and intellectual disability at Z Special Support School. X utilizes a school-provided Chromebook for various
activities both at school and home. However, due to involuntary movements, she encounters difficulties like unintended taps on the screen or unintentional dragging while using the device.

2.3 Procedure

Fittings were conducted four times for a total of 15 minutes per session. The effectiveness of the fitting and its effective method will be evaluated by analyzing the videos taken of the fitting and the episodes from interviews with X after the fitting.

2.4 Ethical Consideration

The research involved in this presentation has been conducted with appropriate ethical considerations and handling of personal information, and consent has been obtained from the individual, parents/guardians, and the head of the school to which the subject belongs.

3. Results

3.1 Fitting #1

During the session 1, X was interviewed about her challenges with using the Chromebook and her interactions with the device were observed. In a previous instance, when submitting an assignment on Google Classroom, X encountered difficulties tapping the touch display accurately, resulting in multiple submission attempts. In a subsequent task, she attempted to use the touchpad for clicking, occasionally applying excessive pressure. Additionally, she demonstrated the ability to search the internet using a combination of touch and keyboard inputs, although the small size of search elements caused her to often tap the wrong areas, leading to frustration. Lastly, updating an "Open Diary" created with Google Spreadsheet proved troublesome for X, as the small columns for typing comments and selecting dates presented tapping challenges.

3.2 Fitting #2

During the second session, efforts were made to address X's struggle with display size. After testing various sizes, a 110% display was selected as the most suitable for her. To address difficulties in clicking, the "auto click" function was introduced, with a one-second delay between clicks to accommodate X's preference and involuntary movements. The pointer position correction function was also implemented. Adjustments were made to the mouse pointer's size, color, and speed based on X's input. The pointer size remained default, the color was changed to red (her favorite), and the default speed was retained for comfortable use. Following these adjustments, it was decided that X would continue using the customized settings until the next session.

*Fig. 1* Using a touch pad alone.  
*Figure 2.* Adjusting functions with teachers.
3.3 Fitting #3

During the third session, an expert in Chromebook manufacturing and sales provided online professional guidance for further fitting adjustments. X shared her concerns about image searches, where her fingers struggled to move accurately, leading to incorrect clicks. The expert taught her to use the tab key for selection, aiding her in choosing the desired image. Techniques for easily navigating selections, even with paralysis, were demonstrated using the Shift and tab keys. X also received guidance on typing comments in the "Open Diary." The expert advised using the adjacent menu column to simplify sheet selection for changing dates, which X found preferable over horizontal scrolling. After the fitting, X consulted the specialist, who taught her sheet-switching shortcuts. However, due to her palsy, she found it challenging to press multiple keys simultaneously and opted against using shortcuts.

3.4 Fitting #4

In the fourth session, X expressed her satisfaction with the customized Chromebook, finding it much easier to use than before. She highlighted the convenience of the physical keyboard, the one-handed display adjustment, and the option to rest her hand on the display's bezel while touching it. As a result, she intends to continue using the Chromebook.

Figure 3. Online advice from an advisor.  Figure 4. Tapping with a finger on the display

4. Discussion

By addressing X's issues and observing her use of the Chromebook, effective fitting was achieved. Altering the display size improved tapping accuracy, while the automatic click function aided clicking despite finger paralysis. Adjusting the mouse pointer's color and speed, based on X's preferences, not only enhanced usability but also fostered attachment to the device. Expert guidance on tab and shortcut keys resolved challenges not solvable by accessibility features alone. The fitting process's success stemmed from understanding the children's problems, adjusting equipment based on their input, and seeking expert advice for issues beyond accessibility improvements.

References


Jirapipat THANYAPHONGPHAT\textsuperscript{a}, Preeyada TAPINGKAEB, Kannika DAUNGCHARONE\textsuperscript{a}, & Krittawaya THONGKOO\textsuperscript{a}

\textsuperscript{a}College of Arts, Media and Technology, Chiang Mai University, Thailand
\textsuperscript{b}Ban San Pa Sak School Chiang Mai, Thailand

\textsuperscript{*}jirapipat.than@cmu.ac.th

Abstract: Motivation has gained considerable attention as educators and researchers seek to understand and nurture the diverse set of 21st century skills required for individuals to thrive in the rapidly evolving landscape of the 21st century. These skills, including critical thinking, collaboration, problem-solving, creativity, and self-direction, among others, are crucial for success in various professional and personal contexts. This study investigates the relationship between 21st-century skills and motivation within the framework of Contextual Inquiry Project-based Learning (CI-PBL). The purpose of the study is to gain insights into how motivation influences the development of various 21st-century skills. The research employs Pearson's correlation coefficient to analyze the relationship between motivation and these skills. The study aims to provide a comprehensive understanding of the strength and direction of the correlations, revealing potential patterns and associations between motivation and skill development.

Keywords: Contextual Inquiry Project-based Learning, Project-based learning, Motivation, 21st Century Skills, ICT innovation in education

1. Introduction

In the rapidly evolving landscape of education, there is a growing recognition of the need to equip students with skills that are relevant to the demands of the 21st century. These skills, often referred to as 21st century skills, encompass a wide range of abilities and competencies that enable individuals to navigate and thrive in the modern world. Some of the key skills emphasized in this context include critical thinking, creativity, collaboration, problem-solving, self-direction, meaningful learning with Information and communication technology, and knowledge creation efficacy.

Individuals’ levels of interest, perseverance, and effort in learning and using new skills are greatly influenced by their motivation. It is well known that it plays a crucial role in educational settings since it affects students’ capacity to set objectives, control their learning processes, and deal with difficulties. Despite the fact that the importance of motivation in the classroom is generally known, research on the precise connection between motivation and 21st century abilities is still underway.

Explore the 21st Century Skills and competencies that are essential for learners in Asia to succeed in the new millennium. The rapid changes brought about by globalization and technological advancement require learners to possess a new set of skills and competencies that go beyond traditional academic knowledge and skills.
Through a study utilizing contextual inquiry project-based learning, this research seeks to analyze this relationship. The study will look at how this strategy can help students develop both of these skill sets more effectively. We intend to learn more about the connections between these two sets of talents and how they might be effectively cultivated using cutting-edge teaching strategies via the course of this research. The results of this study could have a big impact on how educators and policymakers create educational programs that support the growth of computational thinking and 21st century abilities. We can prepare pupils for life in a complicated, technologically advanced environment by giving them these abilities.

The investigation of this relationship study aims to contribute to the existing body of research on both motivation and 21st century skills. By uncovering potential correlations and patterns, educators and researchers can gain valuable insights into the dynamics between motivation and skill development, informing instructional practices and curriculum design.

The research questions in this study are as follows:

- What is the nature and strength of the correlation between motivation and various 21st century skills?
- How does motivation influence the development and application of 21st century skills within the CI-PBL framework?

2. Literature Review

2.1 21st Century Skill

21st-century skills refer to a set of competencies that are considered essential for success in the modern world. These skills go beyond traditional academic knowledge and include abilities such as critical thinking, problem-solving, collaboration, communication, creativity, and digital literacy. According to the World Economic Forum, these skills are becoming increasingly important due to factors such as globalization, technological advances, and the changing nature of work (World Economic Forum, 2018).

Several frameworks have been developed to define and organize 21st-century skills. One popular framework is the Partnership for 21st Century Skills (P21) Framework, which identifies four key categories of skills: (1) communication and collaboration, (2) critical thinking and problem-solving, (3) creativity and innovation, and (4) digital literacy (Partnership for 21st Century Skills, 2015).

Research has shown that developing 21st-century skills can lead to a range of positive outcomes for students, including improved academic performance, increased engagement, and better preparation for future careers (Trilling & Fadel, 2009; Griffin, McGaw, & Care, 2018). However, there are also challenges associated with integrating 21st-century skills into traditional educational practices, such as resistance to change, lack of teacher training, and difficulty in assessing these skills (Voogt et al., 2015).

21st Century Skill by Chai Ching Sing, et al. (2015) is a research paper that proposed a framework for assessing students' perceptions of 21st-century learning practices. The framework includes seven dimensions of 21st-century skills:

1. Self-directed learning (SDL): the ability to take responsibility for one's own learning.
2. Meaningful learning with Information and communication technology (MLT) : the ability to use technology as a tool for promoting student-centered, collaborative, and authentic learning experiences that align with the demands of the 21st-century workplace (Jang et al., 2017)
3. Collaborative learning(CoL): the ability to work effectively in teams and share knowledge with others
4. Critical thinking(CriT): the ability to analyze information and make informed decisions
5. Creative thinking(CreT): the ability to generate and develop innovative ideas and solutions
6. Authentic problem-solving (APS): the ability to solve real-world problems using critical and creative thinking skills

7. Knowledge creation efficacy (KCE): the ability to develop new knowledge through collaboration and efficient learning process.

The paper argues that these seven dimensions of 21st-century skills are essential for students to succeed in the modern world and that educators should prioritize the development of these skills in their teaching practices.

2.2 Motivation

Motivation is the driving force that underlies behavior and actions. It refers to the internal and external factors that influence the direction, intensity, and persistence of behavior (Ryan & Deci, 2020). Motivation plays a critical role in learning and achievement, as it can impact students' engagement, effort, and performance.

Researchers have identified several strategies for fostering motivation in educational settings, such as providing meaningful and challenging tasks, promoting autonomy and choice, giving feedback, and creating a supportive learning environment (Brophy, 2021).

The previous study highlights the effectiveness of project-based learning as a strategy for promoting computational thinking skills and motivating students in computer science education. By engaging students in authentic and hands-on programming projects, the approach allows them to apply computational thinking principles in real-world contexts, enhancing their understanding and interest in the subject matter (Thanyaphongphat et al., 2022). There exist six distinct dimensions that encompass the concept of motivation.

1. IM (Intrinsic Motivation): Intrinsic motivation is the natural desire and interest in doing something for its own sake, such as because you enjoy it or are curious about it. It's important to include this factor because it helps show how interested and passionate people may be about the subject. It's a key way to figure out what really drives people to be engaged.

2. PG (Personal Goals): Personal goals are the hopes, dreams, and goals that a person has for their own academic or personal growth. This factor shows how students' motivation is tied to their own goals, which can be different and can drive their hard work and dedication.

3. EM (Extrinsic Motivation): When someone does something for external benefits or to avoid bad things, this is called extrinsic motivation. This can be based on things like grades, awards, or social acceptance. It's important to include this factor to account for the different motivations that can come from outside sources, which can affect how engaged students are.

4. AC (Anxiety about Computer Science Assessment): Worrying about tests in a certain topic, like computer science, can have a big effect on how motivated students are to learn. This dimension focuses on the emotional side of motivation, showing how worry can both help and hurt motivation. It's important to think about worry because it can change how you learn.

5. SD (Self-Determination): Self-determination is a person's sense of being in charge of their own acts and decisions. This factor is very important for figuring out how much control students feel they have over their learning. Self-determined students are more likely to be self-motivated and interested in school.

6. SE (Self-Efficacy): Self-efficacy is a person's belief that they can do a job or reach a goal on their own. It's a very important factor because it directly affects how hard students are willing to work on a topic. High self-efficacy can help students work harder and get more done.

2.3 Project-based learning

Project-based learning has emerged as a promising approach to fostering both 21st century skills and computational thinking. Contextual Inquiry Project-based Learning, in particular, has been shown to be an effective way of promoting these skills. According to Hsu et al.
(2018), this approach involves students in a collaborative project that is based on a real-world problem. The students are guided through the project with the aim of developing 21st century skills and computational thinking. This approach has been shown to be effective in enhancing students’ problem-solving abilities, critical thinking, and collaboration skills (Hsu et al., 2018).

2.4 Contextual Inquiry

Contextual inquiry is a qualitative research method used to understand users’ behaviors and needs in the context of their work or daily lives. The method involves observing and interviewing users in their natural environment and analyzing the data collected to identify patterns and insights. Contextual inquiry is commonly used in user-centered design and can be applied to a variety of contexts, including healthcare, education, and technology.

One of the key features of contextual inquiry is the emphasis on observation and participation. Researchers not only observe users but also participate in their activities to gain a deeper understanding of their experiences. This approach allows researchers to identify implicit needs and behaviors that may not be apparent through traditional surveys or interviews.

Another important aspect of contextual inquiry is the use of contextual data to inform the design process. Researchers collect data on the physical environment, tools, and artifacts used by users, as well as their social and organizational contexts. This information can be used to develop design solutions that are tailored to users’ needs and contexts.

2.5 Contextual Inquiry Project-based Learning

Contextual Inquiry Project-based Learning (CI-PBL) is an instructional approach that combines the Contextual Inquiry research method with Project-based Learning (PBL) to promote meaningful and authentic learning experiences. This approach involves having students engage in inquiry-based projects that are grounded in real-world contexts and that involve the use of technology. These steps are seven steps for learning:

1. Motivate and Prepare: This step involves building motivation and preparing students with the foundational knowledge and necessary skills for the project.
2. Analyze: In this step, students analyze the problem and make connections to the context in which the project is taking place.
3. Plan: This step involves planning the project and determining how the problem will be solved.
4. Do: In this step, students implement their plan by taking action and solving the problem.
5. Examine & Evaluation: In this step, students examine and evaluate their work to determine its effectiveness and to identify areas for improvement.
6. Explore & Expand: This step involves exploring new knowledge and expanding the scope of the project beyond the original problem.
7. Develop & Distribute: In the final step, students develop and distribute their work to others, sharing their solutions and knowledge with a wider audience.

By following these seven steps, students can engage in authentic, inquiry-based learning that helps them develop critical thinking, problem-solving, communication, and collaboration skills. The CI-PBL approach allows students to take ownership of their learning and connect academic concepts to real-world contexts, making learning more meaningful and relevant to their lives.

Research has shown in Figure 1 that the CI-PBL approach can be effective in promoting the development of 21st-century skills such as critical thinking, creativity, communication, and collaboration. For example, a study by Lee and Wu (2019) found that using a Contextual Inquiry Project-based Learning approach in a college-level computer science course led to improvements in students’ critical thinking and problem-solving skills.
2.6 Relationship between 21st-Century Skills and Motivation in PBL using Contextual Inquiry

There is growing evidence that the development of 21st-century skills is positively related to students’ motivation to learn. Research suggests that when students are engaged in tasks that require them to use 21st-century skills such as critical thinking, creativity, and collaboration, they are more likely to be intrinsically motivated and engaged in their learning (Kirschner & van Merriënboer, 2013).

In addition, project-based learning (PBL) is an instructional approach that is often used to develop 21st-century skills and foster student motivation. PBL involves students working on real-world problems or challenges and requires them to use a range of 21st-century skills such as communication, collaboration, and problem-solving (Hung et al., 2008). Studies have found that PBL can increase student motivation, engagement, and achievement (Krajcik et al., 2014).

3. Methodology

3.1 Participants

The students from northern Thai high schools (grades 7) participated in the pilot project. The same teachers taught the same unit content for groups. This class had twenty-five combined students: thirteen males and twelve females.

3.2 Instrument

The instrument used to assess student motivation was modified from the questionnaire developed by Panjaburee and Sriraswasdi (2016), which had a scale reliability of 0.89. The estimated value of 29 items on Likert’s Scale (Likert’s Scale). The questionnaire scores were allocated a five-level Rating Scale: least, less, moderate, very, and most, which was then translated into Thai by a language expert. Then, retest the confidence with a Cronbach’s alpha confidence of 0.928, which indicates extremely high confidence.

The 21st Century Skills Questionnaire was made by Chai Ching Sing, et al(2015) with a Cronbach’s alpha confidence score of 0.95. We then translated it into Thai. There were 42 questions on the poll, six for each scale. The poll questions were shown on a five-point Likert scale, with 1 being strongly disagree and 5 being strongly agree. The Thai version was used, and Cronbach’s alpha was 0.90, which means there was a very high level of confidence and that it could be used to evaluate 21st Century Skills reasoning in seven different areas.
An open-ended questionnaire focused on the activities that students enjoyed and the experiences that left a significant impression on them.

3.3 Research Design

To investigate the relationship between 21st century skills and motivation skills. The Contextual Inquiry Project-based Learning approach is combined with the context-based inquiry-based learning management to promote 21st-century skills and enhance the motivation of grade 7 students to learn science subjects, specifically Computational Science. The program consists of Level 1, with 1 hour per session, totaling 18 hours, and includes the following units:

- Unit 1: Digital Intelligence (2 hours)
- Unit 2: Digital Problem Solver (2 hours)
- Unit 3: Digital Inventor (10 hours)
- Unit 4: Digital Data Management (4 hours)

The learning activities are designed to integrate skills, processes, and experiences. The content is arranged in order to increase difficulty, catering to the original skills of the learners. Both individual and group activities are included. For group activities, students will be divided into groups of 3-4 people and work together on activity sheets.

The Learning Management Plan for Unit 3, “Digital Inventor,” consists of a 10-hour session for students, with 1 hour dedicated to problem-solving steps. The breakdown of activities is as follows:

- Coding Heart Rate: 1 hour
- Car Coding: 2 hours
- Fan Coding: 3 hours
- Digital Inventor: 3 hours

To support the learning process, various materials will be utilized, including activity sheets, knowledge sheets, slides, and teaching media. Additionally, each student will be provided with one Micro:bit board.

Figure 2. Example Car Coding activity in Digital Inventor unit.

4. Results and Discussions

In order to determine the impact of CI-PBL on Computer Science education, Table 1 displays the level of student motivation after learning. It asserts that the majority of student
learning incentives were positive. In contrast, the evaluation component generated the least amount of anxiety. It indicates that students are not anxious about the test. These results demonstrate that CI-PBL can motivate all students to study Computer Science.

Table 1 Means and SD of the Students’ Motivation Towards the Contextual Inquiry Project-based Learning in Computer Science.

<table>
<thead>
<tr>
<th>Learning Motivations Dimensions</th>
<th>Mean</th>
<th>SD.</th>
<th>SE.</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM (Intrinsic Motivation)</td>
<td>3.7920</td>
<td>0.61232</td>
<td>0.12246</td>
<td>Agree</td>
</tr>
<tr>
<td>PG (Personal Goals)</td>
<td>3.6960</td>
<td>0.75082</td>
<td>0.15016</td>
<td>Agree</td>
</tr>
<tr>
<td>EM (Extrinsic Motivation)</td>
<td>3.5840</td>
<td>0.72093</td>
<td>0.14419</td>
<td>Agree</td>
</tr>
<tr>
<td>AC (Anxiety about Computer Science Assessment)</td>
<td>2.1360</td>
<td>0.77399</td>
<td>0.15480</td>
<td>Disagree</td>
</tr>
<tr>
<td>SD (Self-Determination)</td>
<td>3.6720</td>
<td>0.59127</td>
<td>0.11825</td>
<td>Agree</td>
</tr>
<tr>
<td>SE (Self-Efficacy)</td>
<td>3.5120</td>
<td>0.57178</td>
<td>0.11436</td>
<td>Agree</td>
</tr>
</tbody>
</table>

For another query data analysis, Pearson’s correlation coefficient was used to examine the relationship between the six motivational dimensions. From Table 2, it was shown that the highest positive correlation existed between the intrinsic (IM) dimension and the self-determination (SD) dimension significantly ($r = 0.867$, $p < 0.01$) means that students participate in learning activities with intrinsic motivation and self-determination motivation. They can complete their projects in other ways. Both motives can increase interest and enjoyment in learning. Other significant associations were found between anxiety about the computer science (AC) assessment and self-determination (SD) dimensions. However, the fear of failing the computer science exam affected their decision-making.

Table 2 Pearson’s correlation coefficient between motivation and twenty-first century motivation.

<table>
<thead>
<tr>
<th>Learning Motivations Dimensions</th>
<th>IM</th>
<th>PG</th>
<th>EM</th>
<th>AC</th>
<th>SD</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM (Intrinsic Motivation)</td>
<td>1</td>
<td>0.796**</td>
<td>0.532**</td>
<td>0.319</td>
<td>0.867**</td>
<td>0.674</td>
</tr>
<tr>
<td>PG (Personal Goals)</td>
<td>0.796**</td>
<td>1</td>
<td>0.797**</td>
<td>0.174**</td>
<td>0.798**</td>
<td>0.661**</td>
</tr>
<tr>
<td>EM (Extrinsic Motivation)</td>
<td>0.532**</td>
<td>0.797**</td>
<td>1</td>
<td>0.118</td>
<td>0.636**</td>
<td>0.655**</td>
</tr>
<tr>
<td>AC (Anxiety about Computer Science Assessment)</td>
<td>0.319</td>
<td>0.174</td>
<td>0.118</td>
<td>1</td>
<td>0.397*</td>
<td>0.288</td>
</tr>
<tr>
<td>SD (Self-Determination)</td>
<td>0.867**</td>
<td>0.798**</td>
<td>0.636**</td>
<td>0.397*</td>
<td>1</td>
<td>0.695**</td>
</tr>
<tr>
<td>SE (Self-Efficacy)</td>
<td>0.674**</td>
<td>0.661**</td>
<td>0.655**</td>
<td>0.288</td>
<td>0.695**</td>
<td>1**</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

According to the findings presented in Table 3, the implementation of the learning model resulted in significant improvements in the 21st century skills of the student sample group. Specifically, there was a notable enhancement across five key aspects of these skills: self-directed learning (SDL), collaborative learning (CoL), meaningful learning with ICT (MLT), critical thinking (CriT), creative thinking (CreT), authentic problem-solving (APS), and knowledge creation efficacy (KCE). All of these skills showed a statistically significant increase at the .01 level. Notably, meaningful learning with ICT skills exhibited the highest average score of 4.0240, showcasing the notable impact of this learning approach.
In Table 3, Paired-samples t-test analysis Twenty-first-century skills comparing pre-test and post-test scores.

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDL</td>
<td>3.0080</td>
<td>4.0080</td>
</tr>
<tr>
<td>MLT</td>
<td>2.8560</td>
<td>4.0240</td>
</tr>
<tr>
<td>CoL</td>
<td>3.3520</td>
<td>4.0080</td>
</tr>
<tr>
<td>CriT</td>
<td>3.0400</td>
<td>3.9600</td>
</tr>
<tr>
<td>CreT</td>
<td>3.1700</td>
<td>3.7400</td>
</tr>
<tr>
<td>APS</td>
<td>2.7440</td>
<td>3.9040</td>
</tr>
<tr>
<td>KCE</td>
<td>2.7760</td>
<td>3.8240</td>
</tr>
</tbody>
</table>

**p < .01  

In Table 4, Intrinsic Motivation is positively correlated with most of the twenty-first century skills. Positive correlations indicate that as motivation increases, the corresponding twenty-first century skills also tend to increase. Motivation has a positive and significant correlation with MLT (0.638), CriT (0.775), CreT (0.679), APS (0.265), SDL (0.443), and KCE (0.388).

In contrast, Motivation shows weaker or non-significant correlations with Social and Emotional Learning (SEL) and Active Citizenship (AC). The correlation coefficients for these skills range from -0.223 to 0.126, indicating a weaker or negligible relationship between motivation and these particular skills.

Table 4 Pearson’s correlation coefficient between motivation and twenty-first century skills.

<table>
<thead>
<tr>
<th></th>
<th>SDL</th>
<th>MLT</th>
<th>CoL</th>
<th>CriT</th>
<th>CreT</th>
<th>APS</th>
<th>KCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM</td>
<td>0.443**</td>
<td>0.638**</td>
<td>0.773</td>
<td>0.775**</td>
<td>0.679**</td>
<td>0.265**</td>
<td>0.388*</td>
</tr>
<tr>
<td>PG</td>
<td>0.430</td>
<td>0.691**</td>
<td>0.801</td>
<td>0.732**</td>
<td>0.778**</td>
<td>0.502**</td>
<td>0.591*</td>
</tr>
<tr>
<td>EM</td>
<td>0.404**</td>
<td>0.772</td>
<td>0.725</td>
<td>0.602**</td>
<td>0.623**</td>
<td>0.659*</td>
<td>0.701**</td>
</tr>
<tr>
<td>AC</td>
<td>-0.223</td>
<td>0.085</td>
<td>0.081</td>
<td>0.015</td>
<td>0.018</td>
<td>0.065</td>
<td>0.126</td>
</tr>
<tr>
<td>SD</td>
<td>0.422**</td>
<td>0.645**</td>
<td>0.642*</td>
<td>0.593</td>
<td>0.633**</td>
<td>0.443**</td>
<td>0.516*</td>
</tr>
<tr>
<td>SE</td>
<td>0.377**</td>
<td>0.653**</td>
<td>0.672</td>
<td>0.681**</td>
<td>0.672</td>
<td>0.630*</td>
<td>0.666*</td>
</tr>
</tbody>
</table>

**p < .01  

There are positive correlations between most of the twenty-first century skills factors. Positive correlations indicate that as one skill factor increases, the corresponding skill factors also tend to increase. The highest positive correlations are observed between several pairs of skills factors, indicating strong associations between them. For example, there are strong positive correlations between CoL and CriT (0.820), CoL and MLT (0.769), CoL and CreT(0.697), CoL and SDL (0.493), CoL and APS (0.433), and Collaboration and Knowledge creation efficacy (KCE) (0.578). These correlations range from 0.697 to 0.820,
suggesting that these skills factors tend to be positively related and may complement each other in practice.

Table 5 Pearson's correlation coefficient among the seven twenty-first century skills factors.

<table>
<thead>
<tr>
<th></th>
<th>SDL</th>
<th>MLT</th>
<th>CoL</th>
<th>CriT</th>
<th>CreT</th>
<th>APS</th>
<th>KCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDL</td>
<td>1**</td>
<td>0.518*</td>
<td>0.493</td>
<td>0.648*</td>
<td>0.475</td>
<td>0.315*</td>
<td>0.556</td>
</tr>
<tr>
<td>MLT</td>
<td>0.518**</td>
<td>1**</td>
<td>0.769</td>
<td>0.598**</td>
<td>0.460**</td>
<td>0.638*</td>
<td>0.643**</td>
</tr>
<tr>
<td>CoL</td>
<td>0.493**</td>
<td>0.769**</td>
<td>1</td>
<td>0.820**</td>
<td>0.697**</td>
<td>0.433*</td>
<td>0.578*</td>
</tr>
<tr>
<td>CriT</td>
<td>0.648**</td>
<td>0.598**</td>
<td>0.820</td>
<td>1**</td>
<td>0.732**</td>
<td>0.434**</td>
<td>0.584*</td>
</tr>
<tr>
<td>CreT</td>
<td>0.475**</td>
<td>0.460**</td>
<td>0.697</td>
<td>0.732**</td>
<td>1**</td>
<td>0.245*</td>
<td>0.547</td>
</tr>
<tr>
<td>APS</td>
<td>0.315*</td>
<td>0.638**</td>
<td>0.433</td>
<td>0.434*</td>
<td>0.245**</td>
<td>1</td>
<td>0.743*</td>
</tr>
<tr>
<td>KCE</td>
<td>0.556**</td>
<td>0.643**</td>
<td>0.578</td>
<td>0.584**</td>
<td>0.547*</td>
<td>0.743</td>
<td>1**</td>
</tr>
</tbody>
</table>

Furthermore, the students were engaged in an open-ended questionnaire that specifically centered on identifying activities that they derived pleasure from. The results of the study revealed that students placed a high level of importance on engaging in group activities, participating in coding exercises that involved the utilization of Micro:bit boards, and engaging in programming tasks. They cherished opportunities to initiate projects and independently produce work, particularly in activities such as Fan Coding and Car Coding. Students appreciated the chance to present their personal work, especially in the Fan Coding activity where they were able to program a Micro:bit board to control a fan based on their design specifications and create fan models reflecting their innovative concepts. In terms of experiences that left a lasting impression on the students, it was discovered that students were struck by the teacher’s support during activities, the ability to independently troubleshoot problems, and the collaboration with peers. Students articulated, “Initially, I was unable to complete the task, but after acquiring knowledge and making an effort, I felt a sense of pride and enjoyment.” This feedback illustrates that such activities can enhance motivation for learning and cultivate competencies for the 21st century, specifically in the realms of cooperative and self-directed learning.

5. Conclusion

Research has shown that Contextual Inquiry Project-based Learning can be effective in promoting the development of 21st-century skills such as critical thinking, collaboration, and communication. Similarly, a previous study by Chen and Chang (2016) found that students who participated in a Contextual Inquiry Project-based Learning project showed significant improvements in their critical thinking skills. This aligns with the existing body of research on the utilization of suitable digital media, which has been shown to enhance student motivation and learning efficiency (Thanyaphongphat, 2019).

These results only show associations between motivation and twenty-first century skills, but they do not indicate that one directly causes the other. Furthermore, other factors not considered in this analysis may also contribute to the development of these skills. Future research projects could commence by analyzing student behavior through observations and conducting in-depth interviews to gather qualitative insights. Employing this approach would provide a comprehensive understanding of the situation. In subsequent studies, robust statistical tests could be employed to validate findings, leveraging a larger participant pool. Additionally, it is recommended that future investigations explore the
adaptability of the Contextual Inquiry Project-based Learning method across various digital learning contexts. These aspects share commonalities and possess the potential to contribute to our existing knowledge.

References


Data-Driven Competency Assessment Supporting System for Teachers
Taito KANOa, Izumi HORIKOSHi, Kento KOIKEb & Hiroaki OGATAb
aGraduate School of Informatics, Kyoto University, Japan
bAcademic Center for Computing and Media Studies, Kyoto University, Japan
*kanotaito.87a@st.kyoto-u.ac.jp

Abstract: As many countries seek to promote competency-based education, formative assessments are important to capture the learning processes of learners. However, as yet there are no assessments that can fully capture the learning process. Recently, the use of ICT tools for learning has become more general, and learning log data has been accumulated. Using these data, it has become possible to capture learning processes in detail; therefore, data-driven assessment has attracted increasing attention. However, as conventional data-driven competency assessments require experts to map data to competencies, they can only be applied in a defined context. In this study, we proposed an assessment framework that allows teachers to assess their students’ competency by freely combining data collected as students used the Learning & Evidence Analytics Framework (LEAF) platform. We created an assessment in a scenario in an assumed educational setting using the proposed framework and examined what kind of assessment would be possible. Then, we created a system for the framework. Finally, interviews were conducted with three teachers regarding the system. The results suggest that the system can achieve context-independent and flexible data-driven assessment, contributing to the continuous improvement of learning and teaching from multiple perspectives in activities that use the system.

Keywords: Competency assessment, data-driven assessment, learning analytics

1. Introduction

Competency-based learning, which emphasizes the development of competencies in K12 education, has been widely implemented (Henri et al., 2017). To develop learners’ competencies, it is important to assess their current competencies during the learning process (Guerrero-Roldán & Noguera, 2018). In doing so, it is possible to assess how they have demonstrated and developed their competencies. Competency assessments have also been implemented in K-12 education in Japan (MEXT, 2018). However, most assessment methods focus on products such as tests, reflection sheets, and portfolios; process-based assessments have not yet been developed (Guerrero-Roldán & Noguera, 2018).

Recently, ICT tools for learning have been introduced into the educational field, allowing learners’ activities to be recorded in a database as log data that facilitate capturing the learning process in more detail than ever before (Sung et al., 2016; Van der Kleij et al., 2015). Therefore, data-driven competency assessment has attracted considerable attention.

However, in conventional data-driven competency assessments, experts must define the correspondence between data and competencies theoretically (Greene & Azevedo, 2010; Winne & Perry, 2000). Although this approach achieves a robust and highly valid data-driven competency assessment, it can only assess data as defined and cannot respond to diverse contexts.

Therefore, we created a framework that allows teachers themselves to assess competencies by combining data and implementing a data-driven competency assessment system. We believe that this system will also afford each teacher the flexibility to create assessments tailored to their context and improve the quality of observation. We addressed these RQs in this study.

RQ1: How can the data-driven assessment be realized from trace data?
RQ2: How can the data-driven assessment be implemented for teachers’ use?
RQ3: How do the teachers expect, use and evaluate the data-driven assessment system?

2. Literature Review

2.1 Data-Driven Competency Assessment

To assess competencies such as critical thinking and self-regulated skills (SRSs), it is necessary to capture the learning process (Guerrero-Roldán & Noguera, 2018). Although many methods have been proposed to capture the learning process, they have certain limitations. For instance, assessments based on observations of daily learning activities are limited to what teachers can capture (Greene, 2015). Even if the teacher asks learners to write a reflection sheet, there is still much noise in the sheet, such as the learners’ ability to express themselves and remember (Renninger & Bachrach, 2015). Among the many methods available, assessment based on learning log data has attracted attention as appropriate for assessing the learning process (Siadaty et al., 2016). To assess competencies from the data, the following procedure is necessary.

First, target competency must be defined. Second, the competencies must be subdivided into activity levels. For example, SRSs are divided by Greene and Azevedo (2009) into “macro-level” and “micro-level” activity phases. Third, we must determine the activities in the system that correspond to the activity phase definitions. For instance, if the target competency is SRSs, they are subdivided into “Planning,” “Engagement,” and “Evaluation & Reflection”; “Engagement” is subdivided into “Working on the Task” and “Applying Strategy Changes”; and “Working on the Task” is described as “to consistently engage with a learning task using tactics and strategies” (Siadaty et al., 2016, p. 192). Based on the description, it can be considered to correspond to an activity such as using a marker function to highlight an important word in “Working on the task”. In this case, the system activity must be recorded as data for assessing the activity. By linking activities and data in this way, the assessment value of activities can be quantified from the data, from which the competency can be quantified.

However, such assessment processes are mostly conducted by experts and researchers, and their applicable context is limited. Essentially, the axis of competency assessment, interpretation of data, and context of assessment can be changed by the teacher. In this study, we propose a method that allows teachers to flexibly create assessments according to their own contexts.

2.2 Competency Assessment in Japan

Japan’s K12 program uniquely fosters three pillars of competencies as shown in Table 1 with their descriptions which are considered based on MEXT (2018). Each pillar consists of multiple competencies that are mainly assessed based on learning products such as tests, presentations, and interview tests.

<table>
<thead>
<tr>
<th>Pillar of Competency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge &amp; Skills</td>
<td>How much knowledge and skills learners have acquired</td>
</tr>
<tr>
<td>Thinking, Judgment &amp; Expression</td>
<td>How much acquired knowledge and skills learners can use</td>
</tr>
<tr>
<td>Meta-Cognitive &amp; Self-Regulated Skills</td>
<td>How well learners can recognize their state and adjust their learning to acquire knowledge and skills</td>
</tr>
</tbody>
</table>

The recent global digitalization of education has also progressed in Japan. All learners have begun using devices in accordance with the digital education reform known as the GIGA School Concept (MEXT, 2020). There are high expectations regarding the use of learning-log data obtained from ICT tools for competency assessment. However, the use of learning-log data for assessment has not yet been fully realized.
3. Method

3.1 Learning & Evidence Analytics Framework (LEAF)

The LEAF system that our laboratory is developing is an integrated learning platform that integrates LMS (Moodle), e-books (BookRoll), a learning support module, a learning analytics module (LogPalette), and a database (LRS) (Ogata et al., 2018). LEAF can serve as a sensor that records learners’ detailed learning behaviors and a database that stores learning logs. The following table presents example BookRoll log data. In this study, a data-driven assessment system was created using the learning-log data obtained from LEAF. However, this assessment method can be applied in other environments as well.

<table>
<thead>
<tr>
<th>Operation_time</th>
<th>Student_id</th>
<th>Course_id</th>
<th>Operation_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022-04-05 13:32:01</td>
<td>S_1</td>
<td>C_1</td>
<td>“ADD_MARKER”</td>
</tr>
<tr>
<td>2022-04-06 07:14:43</td>
<td>S_2</td>
<td>C_1</td>
<td>“OPEN”</td>
</tr>
<tr>
<td>2022-04-06 08:01:02</td>
<td>S_2</td>
<td>C_1</td>
<td>“NEXT”</td>
</tr>
</tbody>
</table>

3.2 Research Questions

- **RQ1: How can the data-driven assessment be realized from trace data?**

  We began by creating an assessment framework that allows teachers to combine data and create assessments. As an example application of this framework, we attempted to assess learners’ SRSs in a first-year junior high school mathematics course using a set of created indicators. The results indicated the type of assessment possible using the framework.

- **RQ2: How can the data-driven assessment be implemented for teachers’ use?**

  This section describes the implementation of this assessment. In addition, we describe some of the support functions added to assist teachers in the assessment.

- **RQ3: How do the teachers expect, use and evaluate the data-driven assessment system?**

  We conducted interviews with three Japanese teachers who used the implemented system. We compared it with the current competency assessment and asked participants about their expectations and concerns. We then asked them to use the system, and based on observations of their behavior, we discussed its positive aspects and possible improvements.

4. Results and Discussions

4.1 RQ1: Data-Driven Assessment Framework

Here, the proposed assessment framework is introduced and applied to a given scenario.

4.1.1 Indicators and Data Processing Flow

The following explains the process for calculating the assessment value using trace log data, as shown in Figure 1.

![Figure 1. Data Processing Flow](image-url)
First, the indicators were created by aggregating the trace log data obtained from the LEAF. Second, the indicators were narrowed down by the time range of the activity to be assessed, and summed. Next, the logarithmic transformation of each indicator across learners was performed to make the distribution closer to a normal distribution, and standardization was performed to align the distribution means of different indicators. In this way, the influence of outliers can be suppressed, and different indicators can be summed. Finally, by selecting the indicators to be used for assessment and setting weights for each, the values of the selected indicators were weighted and summed to calculate the assessment value. In conventional assessment activities, the scores on several tests or the results of a presentation are weighted and summed to create a rating. Based on these suggestions, ratings were created by adding weights to the indicators. In addition, summative assessments can be created by aggregating the indicators over the entire period, and formative assessments by aggregating the indicators for each week.

This procedure was expected to enable not only the generation of learner ratings from the data but also continuous monitoring of the changes and growth of learners. Nine indicators were created in this study, as listed in Table 3.

<table>
<thead>
<tr>
<th>Indicator Name</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>Time Spent</td>
</tr>
<tr>
<td>NT</td>
<td>Number of Trans</td>
</tr>
<tr>
<td>NJ</td>
<td>Number of Jumps</td>
</tr>
<tr>
<td>NYM</td>
<td>Number of Yellow Markers</td>
</tr>
<tr>
<td>NRM</td>
<td>Number of Red Markers</td>
</tr>
<tr>
<td>NB</td>
<td>Number of Bookmarks</td>
</tr>
<tr>
<td>NM</td>
<td>Number of Memos</td>
</tr>
<tr>
<td>NHM</td>
<td>Number of Hand Writing Memos</td>
</tr>
<tr>
<td>NAQ</td>
<td>Number of Attempts of Quiz</td>
</tr>
</tbody>
</table>

**4.1.2 Scenario Analysis**

To demonstrate the application of the framework, we conducted an assessment with these nine indicators in a scenario that assumed an actual educational setting as follows:

A teacher teaching a first-grade math course in 2022 will assess his or her students’ first semester SRSs. This course will also have a test on June 3. Test scores will be used to assess Knowledge and Skills; therefore, learning activities in the LEAF from April 1 to June 3 will be assessed as a component of SRSs. Figure 2 shows the results of the assessments of the SRSs of the three students based on the indicators shown in Table 3 using the assessment process shown in Figure 1.

Figure 2(a) shows the value of the assessment indicators calculated for the three students during the entire period. Red and blue lines are drawn for indicators in the top and bottom 30% of the values, respectively. These show that student_1 scored high only in the number of red markers, student_2 in four indicators, and student_3 in five indicators. Each student varied in whether indicators were high or low, which may indicate differences in students’ approaches. We believe that these figures are important for understanding the characteristics of learning in LEAF enhanced learning activities.

Second, we considered the actual assessment of SRSs based on these nine indicators. Among SRSs, we will focus only on Working on the Task activities for the assessment target because the indicators are considered to be related to Working on the Task, which is described as “to consistently engage with a learning task using tactics and strategies” (Siadaty et al., 2016, p.192).
The activities performed in Working on the Task were further classified into three categories (Table 4), and each indicator was mapped to a category based on its characteristics.

Table 4. Correspondences between Categories and Indicators

<table>
<thead>
<tr>
<th>Category</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>Viewing time, number of page transitions, number of page jumps</td>
</tr>
<tr>
<td>Applying Strategy</td>
<td>Number of yellow markers, red markers, bookmarks</td>
</tr>
<tr>
<td>Exercise</td>
<td>Number of notes, handwritten notes, quiz responses</td>
</tr>
</tbody>
</table>

The calculated assessment value of each activity was assigned A for the top 30%, B for the top 30–70%, and C for the remaining activities to be graded. The grades for each activity from April 1 to June 3 are listed in Figure 2(b). Test grades are also provided for reference purposes.

The results of this summative assessment provide insights into students’ learning activities that were not found in test scores alone. For example, student_1, who also scored high on the test, did not perform many Reading or Exercise activities but was highly involved in the Apply Strategy activity, suggesting that he or she was trying to learn creatively. Student_3, who has high scores in each activity but a low score on the test, is considered to have some problem with the way he or she performs the activity. These differences in the learning activities of each student afford a perspective for assessment that cannot be seen from grades alone.

Thus far, we have learned how to incorporate the data obtained from LEAF into a conventional summative assessment. Such summative assessments provide suggestions on how to conduct learning activities but do not capture the learning process. Therefore, we calculated the indicators for each week and performed weighted additions to calculate the assessment values for the time series. Figure 2(c) shows the transitions in the assessment values of the three activities for all students.

This formative assessment makes it possible to see how long the activities had continued and when they were decreasing or increasing. For example, student_2 improved on Reading and Apply Strategy activities from the first week of May to the third week of May, but then...
dropped from the fourth week of May. Student_3 improved on Reading and Apply Strategy activities as well from the first week of May, and continued to do so until the test. Using a time-series graph allows us to capture student activities’ real-time changes and also allows us to observe multiple activities simultaneously. We believe that this will allow for a more multifaceted and continuous formative assessment in LEAF enhanced classes.

In other cases, when a teacher makes a specific intervention for students in a LEAF enhanced classroom activity, it is possible to evaluate their intervention based on how the student activities have changed. We believe that this will help activate the cycle of classroom practice and improvement.

To summarize the assessment process, summative and formative assessments were created by selecting indicators from a given set of indicators according to their purpose. Thus, the framework allowed us to flexibly tailor the assessment to each context. Furthermore, this could serve not only as a grading instrument but also to improve learning and teaching in the system.

However, such freedom in assessment may reduce its validity. Similar issues have been raised in previous studies (Azevedo, 2015; Guerrero-Roldán & Noguera, 2018). When assessing competencies based on data, it is necessary to establish a correspondence between the indicators and competencies. This mapping is mostly performed by experts and researchers who are familiar with both data and competencies, as we did here, and is considered difficult for teachers, whose comprehension of competencies and indicators varies widely. However, teachers should be able to perform the mapping to create assessments flexibly according to their own context. That is, it is necessary to consider how teachers can learn to define competencies and make appropriate correspondences. Accordingly, in this study we incorporated a function to store and share the correspondences made by teachers between indicators and competencies in the system.

4.2 RQ2: Implementation of the Proposed Assessment Method

Figure 3 shows the three main modules included in the system. The main components of the system are the Setting Module, which sets the context of the assessment and selects indicators; the Sharing Module, which shares information to support the selection of indicators in the Setting Module; and the Visualization Module, which visualizes the calculated assessment values.

![Figure 3. Main Modules of the System](image)
4.2.1 Setting Module and Sharing Module

The Setting Module is shown in Figure 3(a). It is divided into two sections, one for inputting the context of the assessment and the other for selecting indicators. In the context input section, the name of the activity to be assessed, competency to be assessed, type of activity, and assessment period are input. In the indicator selection section, a combination of indicators can be set to calculate the assessment values by setting a weight of 1 or more for the indicators to be used, and a weight of 0 for the others.

In participatory design systems, such as the Setting Module, where data use is teacher-dependent, teachers’ lack of expertise is often problematic (Dilmore et al., 2013; Holt et al., 2015). Sharing prior knowledge of data use with teachers is said to be effective in addressing this issue (Dollinger et al., 2019). This can complement teachers’ expertise. Furthermore, showing examples of data use by other teachers improves the reliability of these examples. Therefore, in this study, after inputting the context information into the Setting Module, the following information in a similar context was shared by pressing the search button in the Sharing Module (Figure 3(b)).

b-1. The three most recent assessment cases
b-2. The percentage of selected indicators
b-3. A weighted average of indicators when selected
b-4. Estimated weights predicted by the system

With this information, we attempted to bridge the gap between the nature of the indicators and the teachers’ understanding of them. As the assessments created by the project continue to accumulate as case studies, the generalizability of shared information will increase. We believe that this will lead to improvements in the validity of the correspondence between the selected indicators and competencies.

4.2.2 Visualization Module

The Visualization Module, shown in Figure 3(c), provides the visualization of teacher-generated assessments. It is divided into the following two sections

c-1. Visualization of summative assessments
c-2. Visualization of formative assessments

Based on the results obtained in 4-1-2, we considered the requirements for each section. c-1 has two requirements: (i) multiple assessment values should be visible together, and (ii) the average of the indicators over the entire period should be available. First, to achieve (i), we visualized multiple assessment values on the same graph and table so that all assessment values could be grasped at once. Second, to achieve (ii), the average values over the entire assessment period were described in a radar chart and in three levels of ABC, so that the average characteristics of learning over the period could be captured. Achievement of these two requirements is expected to lead to a comprehensive understanding of the characteristics of learning activities.

In c-2, the three requirements were (iii) that multiple assessment values be visible together, (iv) that the process of changing assessment values be visible, and (v) that it be possible to see when the change occurred. To achieve (iii), multiple assessment values were visualized on the same graph as in c-1. To achieve (iv), the assessment values were calculated according to a time series so that the process of change in the assessment values could be seen. Finally, to achieve (v), the assessment values were calculated weekly to show in which week the change occurred. Fulfillment of these three requirements is expected to lead to continuous assessments and measurement of the effectiveness of instruction.

In summary, this module has the functionality to perform a comprehensive and continuous assessment based on multiple assessment values. It is expected to contribute to the observation of the characteristics and changes in learning activities and the effects of their own interventions.
4.3 RQ3: Interview about the System

To investigate their impressions of the developed data-based assessment system, we interviewed three teachers (one math, two English) from high schools that have implemented the LEAF system.

4.3.1 Expectations and Concerns

Each teacher was asked about his expectations and concerns about the system compared to traditional competency assessment. As expected, the English teachers responded that it was possible to assess independent reading activities. The mathematics teacher responded that there were activities that were not fully captured in the field and that the system could be used to assess such activities. He also said that it would be good to be able to assess indicators that can only be qualified by data (e.g., average solving time). One concern raised by teachers in both subject areas was that not all students’ activities were conducted in the LEAF system, and even when they used it, they were not always performing the activities they were supposed to.

This indicates that teachers expect two things from data-driven assessment: that it will allow them to understand what is beyond their reach, and to quantify students’ activities with objective measures. We also found that they were concerned that the activities using the system would not be fully accomplished. Therefore, data from more sensors should be converted into indicators to increase their diversity and promote the activities supported by LEAF.

4.3.2 Experience and Observation

Next, we asked a mathematics teacher to use the system and observe its usage to elicit reflections of teachers. This was conducted shortly after a demonstration of the system. Thus, although he had received a brief explanation of the system’s functions and purpose, he had never directly used the system. In the same context as in Section 4.1.2, we asked him to select indicators for assessing students’ SRSs in exercise activities in two patterns: with and without the use of the function for sharing assessment cases. In both patterns, he selected indicators from a state in which nothing is selected. Figure 4 shows the results of the indicator selection for each case. Indicators not selected for either pattern are not shown.

![Figure 4. Two Patterns of Indicator Selection by the Teacher](image)

Also, while he silently selected indicators from start to finish when he did not use the sharing function, he commented on it as follows.

(u1) “It would be nice to see the title (author's note: name of the indicator), I would like to know the purpose for which the indicator was created.”
(u2) “I guess different subjects have different choices.”
(u3) "Just a quiz would be fine."
(u4) "Even if the others’ selections of indicators are shared, the selection would not be affected very much."

These differences provide some insight into the effect the feature had on him. First, in statement u3, while he considered quizzes important to assess SRSs, he gave less weight to the number of quiz attempts than the other indicators without the sharing function. However,
when using the sharing function, he gave the maximum weight to this indicator. Furthermore, in u4, while he said that the selection was not much affected, the actual selection changed between the condition with sharing function and that without. These findings suggest that the information about others’ choices shared by the sharing function provided him with some awareness of which indicators were actually important to him.

Contrarily, a weak point is that, as stated in u1, the name or purpose of the indicator was not indicated. This implies that to refer to the shared information, it is important that its purpose matches one’s own purpose. In the future, it is necessary to improve the sharing function and share the purpose itself so that teachers can refer to the shared information.

Finally, concerning u2 and u3, some statements focused on the differences in subjects and emphasized the importance of quiz activities that they often use. This suggests that teachers may choose indicators based on their own context and teaching style. Therefore, we will examine whether these factors actually influence teachers’ choice of indicators and also investigate the possible influence of other factors in future research.

5. Conclusion and Future Work

To achieve data-driven competency assessment by teachers, we created a framework that allows teachers to freely select indicators and implemented a system that shares information about other teachers’ assessments. The system differs from conventional data-driven competency assessment systems in that teachers select their indicators by themselves. While this has the advantage of allowing teachers to tailor their assessments to their context, it also decreases the validity of the assessment. However, the accumulation of assessment cases is expected to contribute to addressing this issue. In addition, the interviews with teachers revealed that the system seemed to allow the capture of independent learning and detailed efforts that could not have been captured before. However, it became clear that there were problems outside the system, such as a lack of sufficient activities using the system. Furthermore, when teachers were asked to use the system, it was found that the Sharing Module helped teachers gain awareness regarding indicator selection and that their indicator selection changed. In the future, we would like to increase the effectiveness of the shared information by clarifying what teachers consider important in selecting indicators.

In summary, the proposed system and functionalities are expected to achieve a context-independent and flexible data-driven assessment. Furthermore, it will be possible to continuously improve learning and teaching with activities using the system from multiple perspectives. However, in the specific assessment cases presented in this study, only the logs obtained from LEAF were used for assessment, and there was a limitation that the data obtained from off-system activities and other tools were not included in the assessment target. However, xAPI format-compliant data are available in the system. Therefore, if other tools are also xAPI compliant, cross-data source assessment will be possible in the system. Furthermore, if more sensors record off-system activities as data, the system can be extended to off-system activities. In this way, we hope to make it possible to conduct highly valid assessments in a variety of contexts.

Acknowledgments

This study is supported by NEDO JPNP20006, JSPS KAKENHI 22K20246, and 23H00505.

References


Visualization of Instructional Patterns from Daily Teaching Log Data

Kohei NAKAMURA\(^a\), Izumi HORIKOSHI\(^b\), Rwaitjit MAJUMDAR\(^b\), Hiroaki OGATA\(^b\)
\(^a\) Graduate School of Informatics, Kyoto University, Japan
\(^b\) Academic Center for Computing and Media Studies, Kyoto University, Japan
*nakamura.kohei.42@st.kyoto-u.ac.jp

Abstract: Reflections are said to improve teachers’ professional skills. Questionnaires and video-recording methods require advanced preparation, which is often burdensome for teachers and students. In recent years, researchers have turned to analyzing the log data of ICT systems. This method is less demanding and helps clarify the educational process. Also, with multiple classroom systems, we propose a visualization method to track teachers’ instructional patterns. Our goal is to clarify differences in teaching using actual teacher-log data. It was found that the method of using the system differed depending on each teacher. This method allows us to improve education quality by facilitating teachers’ reflections and fostering discussions with colleagues.

Keywords: Instructional pattern, xAPI, teaching analytics

1. Introduction

Reflection on teaching practice plays an important role in teacher professional development (TPD)(Postholm, 2008). It is not only effective to reflect on one’s educational practices but also on those of other teachers’ educational practices. (Kleinknecht & Schneider, 2013, Koukis, & Jimoyiannis, 2019). TPD encompasses various approaches, but its ultimate objective is to improve practice and facilitate student development (Avalos, 2011). Therefore, it is necessary to gather data from daily classroom practice.

In general class practice, data collection often collected time slots and methods such as questionnaires, audio, and videos (Cherrington & Loveridge, 2014). However, this process can be burdensome for teachers and costly. (Martínez et al., 2022). Therefore, there is growing interest in utilizing teachers’ log data collected from daily classroom practice (Hoyos & Velásquez, 2020). For instance, Reguera et al. (2022) used dashboards to visualize teachers’ utilization of learning management systems (LMS) based on log data to support reflection.

It is important to recognize that teachers often use multiple ICT tools for effective instruction (Mangaroska et al., 2021). Consequently, relying on data from a central platform, such as an LMS, may not accurately the dynamics of a real classroom. Examining how teachers employ multiple systems and analyzing the teaching process is crucial to gain a comprehensive understanding of teaching practices. (Leeuwen et al., 2019) For instance, Nakamura, Horikoshi, and Ogata (2022) utilized xAPI to analyze teaching practice across multiple ICT tools and visualize the process.

The teaching practice and process involve instructional patterns, which are strategic methods teachers used by teachers to guide student learning. Lin et al. (2020) conducted a study on instructional patterns for teaching and learning argumentative writing, where they observed how teachers combine various methods to teach different argumentative components. We aimed to identify the instructional patterns of Japanese junior high school teachers.

In this study, we aim to clarify the difference in teaching by utilizing multiple systems. We will improve Nakamura et al.’s method (2022) and visualize the instructional patterns of teachers. The research questions guiding our study are as follows:

RQ1: What kind of ICT tools do teachers use to teach daily?
RQ2: Do instructional patterns differ depending on tool use experience?
RQ3: Is there a difference in instructional patterns using ICT tools among the same subjects?

2. Method

2.1 Learning Evidence Analytics Framework (LEAF)

The LEAF system (Figure 1) was used to collect Log data for this study. The system comprises three subsystems: the learning management system (LMS) “Moodle,” the e-book reader “Bookroll,” and the learning analysis tool “Logpalette” (Ogata et al., 2018). These subsystems are independent systems interconnected through Learning Tools Interoperability (LTI). The log data from each system were stored as xAPI in the Learning Record Store (LRS). Teachers are in a digital educational environment where LEAF can be used daily, but its use is at the teacher’s discretion.

![Figure 1. Overview of the LEAF system](image)

2.2 Participants and Contexts

This study used observational data from the LRS and involved 22 junior high school teachers. The data were collected from the LEAF system between April 1st and September 30th, 2022. Upon analyzing data from 22 teachers, it was observed that most teachers who used multiple systems daily were math teachers.

2.3 Data parsing and visualization technique

During the study period of 6 months in 2022, a total of 22 teachers generated 53,393 statements. To focus on the use of ICT tools, we filtered logs where these tools were used at least twice. This resulted in the extraction of 3,095 classes based on the timetable. Specifically, when we extracted the classes of four mathematics teachers, we found that there were 552 classes during the target period. From these classes, we further extracted 75 classes to specifically analyze the instructional patterns within the same subjects. We analyzed the xAPI statements and extracted information on (1) the system used, (2) the specific tool used, and (3) the type of operation performed (Figure 2). Through this analysis, we identified a total of 29 types of processed data. Additionally, based on the functions of each system (LMS, E-
book reader, and Learning Analysis tools), as shown in Figure 3, we prepared and analyzed 12 types of teachers’ behaviors.

**Figure 2.** The procedure of data parsing and the number of data points collected in this study.

**Figure 3.** Types of teacher behavior in the LEAF system and their corresponding processed data.
All the classified processed data were point-process data. It is important to explain which tools were used, when, and for how long to gain a better understanding of the instructional process. Figure 4 illustrates this by representing click events as red dots and adding the time between events to depict tool usage time. Additionally, to identify the system from which the data originated, LMS-related data were assigned the color orange, e-book-related data were presented in blue, and analytics tool-related data were shown in green. Each event was assigned a unique color. The y-axis represents the processed and classified data, as shown in Figure 4, while the x-axis represents time. The time is divided into intervals of 1 min, with each interval corresponding to half of the timing, from the start to the end of the lesson. Finally, we visualized the teaching process, as shown in Figure 5.

![Figure 4. How to visualize teachers' activity from point process data](image)

**2.4 Data Analysis Method**

We performed three analyses to answer the three RQs we set.  
(1) **Analysis 1: Exploring the Daily Use of ICT Tools in Teaching (To answer RQ1)**  
During analysis 1, we investigated the systems utilized by the four mathematics teachers used in their classroom practice and compared them to the usage pattern of all 22 teachers. As part of this analysis, we aggregated and analyzed the combination of each system and the number of classes held.  
(2) **Analysis 2: Analyzing Differences in Instructional Patterns Based on LEAF Experiences (To answer RQ2)**  
We overlaid the diagrams generated using the proposed method (Figure 6) with image transparency set to 0.5. Therefore, the overlapped portion appeared in a dark color. By referring to the usage history and usage of ICT tools, we clarify the differences in teachers’ instructions across multiple systems.  
(3) **Analysis 3: The differences in instructional patterns using ICT tools (To answer RQ3)**
To investigate differences in instructional patterns, we examined 75 classes taught by four mathematics teachers within the designated period. The analysis involved using the proposed visualization method to identify and analyze variations in instructional approaches.

3. Results and interpretation

3.1 RQ1: What kind of ICT tools do teachers use to teach daily? (Analysis 1)

In addressing our first research question, we sorted to gain a deeper understanding of the specifics of ICT tool usage. Then we uncovered the preferences and practices of teachers in their daily classes. Figure 7 shows the systems employed in teachers’ classes from April 1st to September 30th, 2022. In the analysis of class situations for the 22 teachers (Figure 7(a)), we found that LMS was the most frequently used ICT tool independently (N = 2411, 77.9%). The second most common combination was LMS and E-book reader (N=507, 16.4%).

Conversely, Figure 7(b) reveals the combination of LMS and E-book reader was the most popular system choice among all four math teachers (N=369, 66.8%). The next frequently used classes were E-book reader (N=73, 13.2%) and classes solely utilizing an LMS (N=65, 11.8%). This indicates that mathematics teachers used the system differently than teachers in charge of other subjects.

(a) Usage by all 22 teachers in 3,095 classes
3.2 RQ2: Do instructional patterns differ depending on tool use experience?

Expanding upon our initial findings, we next explored whether teachers’ experience with the LEAF system influenced their instructional patterns and methodologies. We delved into the profiles of four mathematics teachers at this junior high school. Teacher A, having been with the institution since 2010, assumed the role of teaching 3rd-grade mathematics in 2022. Meanwhile, Teacher B joined the institution and taught 2nd-grade mathematics in 2022. Teacher C has been a part of the institution since 2020, while Teacher D recently joined the institution in 2022, both teaching 1st-grade mathematics in 2022. Regarding their proficiency with the LEAF system, Teacher A stands out with four years of experience, followed by Teacher C with two years. Teacher B has years’ experience, while Teacher D is relatively new, with less than a year of engagement. A comprehensive overview of these teachers’ backgrounds is presented in Table 1.

Table 1. Four mathematics teachers

<table>
<thead>
<tr>
<th>Class</th>
<th>Teacher A</th>
<th>Teacher B</th>
<th>Teacher C</th>
<th>Teacher D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment period</td>
<td>from 2010</td>
<td>from 2021</td>
<td>from 2020</td>
<td>from 2022</td>
</tr>
<tr>
<td>LEAF use experience</td>
<td>4 years</td>
<td>1 year</td>
<td>2 years</td>
<td>Less than 1 year</td>
</tr>
</tbody>
</table>

We aimed to visualize teachers’ instructional patterns from daily teaching log data to understand their experience and familiarity with the LEAF system. We used the overlay method proposed in Figure 6. Figure 8 shows a visualization of all classes conducted during the specified period. In this case, Teacher A, who teaches 3rd grade, used learning analysis tools for a long time and tended to use them in the first half of the class (Figure 8(a)). Moreover, when compared to Figure 8(b), (c), and (d), it is apparent that Teacher A used multiple functions, such as forums and discussions during this period. These results suggest that the use of ICT tools changes as the school year progresses, resulting in the inclusion of diverse instructional methods.
3.3 RQ3: *Is there a difference in instructional patterns using ICT tools among the same subjects?*

In our analysis of teachers utilizing various ICT tools in 75 classes, we analyzed the specific tools they specifically used. The breakdown of these 75 classes is as follows: Teacher A conducted 15 classes, while Teacher B classes conducted 13 classes. Additionally, there were 47 classes taught by teacher C and teacher D combined (Table 2).

Following the analysis of 75 classes, we identified two distinct categories in teachers’ instructional patterns. Firstly, there were classes that solely utilized e-books (N=18, 24.0%). Secondly, there were classes using LMS and e-books (N=57, 76.0%). We then analyzed the differences in instructions using the proposed visualization method.

<table>
<thead>
<tr>
<th>Learning Units</th>
<th>Teacher</th>
<th>Number of classes</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapes and Similarity (Textbook, pp.120-153.)</td>
<td>Teachers A</td>
<td>15</td>
<td>Mid-April - Mid May</td>
</tr>
<tr>
<td>Investigate Shapes (Textbook, pp.94-135.)</td>
<td>Teacher B</td>
<td>13</td>
<td>Mid-May - Mid-June</td>
</tr>
<tr>
<td>Proportion and Shapes (Textbook, pp.91-117.)</td>
<td>Teacher C</td>
<td>20</td>
<td>Early-June - Early September</td>
</tr>
<tr>
<td>Proportion and Shapes (Textbook, pp.91-117.)</td>
<td>Teacher D</td>
<td>27</td>
<td>Early-June - Early September</td>
</tr>
</tbody>
</table>

Figure 9 presents a case study of a class that exclusively used e-books. The Teachers’ usage of e-book can be classified into two categories based on the timing of “textbook” use. The first pattern used the tool all the time during the class (N=14, 77.7%), as shown in Figure 9(a), and the other used the tool half the time (N=4, 22.2%), as shown in Figure 9(b).
In contrast, Figure 10 illustrates the categorization of classes that utilized both LMS and E-books. The analysis reveals that the majority of classes predominantly relied on two tools: “textbook” and “course-viewed” (N = 51, 89.5%), as demonstrated in Figure 10(a). However, a smaller subset, as presented in Figure 10(b), incorporated an additional tool, “Annotation”, resulting in the use of three distinct tools in total (N=6, 10.5%).

Figure 10 Classes using LMS and e-books together (N=57 classes by four teachers) (B2, B5, and B7 indicate the names of teachers’ behaviors shown in Figure 3)

4. Discussion

The study aimed to clarify the differences in class instruction using multiple systems to support teachers’ reflections on teaching practices for TPD. Through our analysis, we identified that the usage of ICT tools differs depending on the subject. However, there was a commonality in tool usage within the same subject. Furthermore, we observed that the usage and timing of ICT tools differed based on the experience and role of the teacher.

Reflection from experience and practice is important to promote behavioral change and decision-making and improve teachers’ professional skills (Prieto et al., 2020). However, in the context of analyzing teachers’ activities, the focus has been on monitoring the students rather than closely observing teachers’ activities (Ndukwe & Daniel, 2020). Therefore, teachers have traditionally relied on questionnaires and videos to reflect on their practice. However, the method proposed in this research offers a novel approach that allows teachers to reflect on their classes without requiring any special preparation using ICT tools daily. This method also facilitates the sharing and discussion of teaching practices with others.
Essentially, this research supports teachers’ reflections without the need for extensive preparation. Three key findings come out of this study.

(1) **RQ1: What kind of ICT tools do teachers use to teach daily?**
Our findings underscore the value of daily log data in understanding teaching practices. While our study did not delve into specific tools, it highlighted the variance in tool usage patterns based on teachers’ backgrounds. It became clear that different teachers used the same system differently.

(2) **RQ2: Do instructional patterns differ depending on tool use experience?**
Research question 2 focused on four mathematics teachers and aimed to identify patterns in how teachers used ICT tools based on their tenure and role. To accomplish this, we employed the overlay method proposed in the study. The results indicated that senior educators, perhaps seasoned by their prolonged interaction with the system, seem to exhibit a broader repertoire in their teaching approach. Meanwhile, newer entrants appeared to be on a learning curve.

(3) **RQ3: Is there a difference in instructional patterns using ICT tools among the same subjects?**
In research question 3, we conducted an analysis of the usage of specific tools in 75 mathematics classes. Our research findings indicate that there are variations in teaching styles with the same subject. We can explore differences in teaching styles from extensive daily log data by using this visualization. By identifying the differences in teaching styles depending on the subject, teachers can focus on important differences in practices and reflect on them.

This study has certain limitations that should be acknowledged. Firstly, the identification of classes was based on rules, dates, and timetables extracted from log data. Additionally, the estimation of instruction types was also based on the log data. Moreover, the results of this study are limited to Japanese junior high school teachers. On the other hand, we have presented a method for visualizing xAPI log data. Therefore, this method has not yet been evaluated by teachers. In the future, we plan to involve teachers in the evaluation of the proposed method.

In the future, we intend to extend the proposed method to include other teaching materials and subjects. Teachers utilize various tools, such as ICT, to deliver the content of their subjects. Therefore, the use of ICT tools may vary depending on the specific content being taught. The proposed method has the potential to extract teachers’ instructional patterns within the same subjects. By using ICT tools, it becomes possible to identify effective teaching methods through log data analysis.

5. **Conclusion**

In this study, we enhanced Nakamura et al.’s (2022) method to support teachers’ reflections on class practice and clarify the differences in class instruction using multiple systems. By applying this method to the xAPI log data of Japanese junior high school teachers, we were able to visualize the tools utilized by the teachers across multiple systems. However, the Proposed visualization method was not evaluated, so we lack insight into how real teachers might use the visualization. In the future, it is necessary to present the results visualized by this proposed method to actual teachers and assess their effectiveness. By employing the proposed method, teachers can reflect on their daily use of ICT tools based on the logged data, which has the potential to improve their professional skills.

**Acknowledgments**

This study was supported by NEDO JPNP20006, JSPS KAKENHI JP22K20246 and JP23H00505.

**References**


Sustaining students’ interest in an instructional system design course by leveraging interest-driven creator theory

Vishwas BADHE\textsuperscript{a}, Sunita RASTE\textsuperscript{a}, Sahana MURTHY\textsuperscript{a}, and Sridhar IYER\textsuperscript{a}

\textsuperscript{a}Indian Institute of Technology, Bombay, India
*vishwasbadhe@iitb.ac.in

Abstract: In semester-long courses, the trajectory of students’ interest hits multiple highs and lows as the course progresses. In general, sustaining students’ interest is a challenge, which becomes even more important when there is a real-world project with an expected deliverable. This study emphasizes investigating how instructors’ strategies help in sustaining students’ interest in instructional system design (ISD) course. The course participants developed a real-world project involving training resources for sensitization about the needs and challenges of persons with disabilities (PwD) in an educational institute. We have analyzed interview data of fifteen students and one instructor, survey data of the same students, and in-class observations by researchers. We have identified instructors’ strategies which were implemented thoughtfully and supported by interest-driven creator (IDC) theory. We have mapped indicators from student responses to the components of the interest cycle i.e. triggering, immersing, and extending. The findings of the study show that course strategies supported by IDC theory played an important role in sustaining the students’ interest throughout the ISD course while ensuring both conceptual learning and project execution.

Keywords: Interest, Instructional systems design (ISD), Interest-driven creator (IDC) theory, project-based learning, persons with disabilities (PwD)

1. Introduction

Design is a process of engaging in a disciplined inquiry for creating something new with practical utility. Instructional design (ID) is a subset of design, directed toward the practical purpose of learning (Rowland, 1993). A typical instructional system design course covers a conceptual understanding of instructional design (ID) models, application of the models, understanding of different aspects of an ID project execution, etc. In instructional design, it’s important that students not only understand the concepts but should also be able to follow and carry out the processes outlined in a model (for example, analysis, design, development, implementation, and evaluation in the ADDIE cycle). To achieve this, many instructional system design (ISD) courses follow a project-based learning approach where students get an opportunity to apply and test their conceptual understanding. In project-based learning, students engage with the problem, learn by doing, discuss, and apply ideas to solve the problem given to them. This increases students’ engagement and helps them to develop a deeper understanding of important ideas by facilitating opportunities for problem-solving, decision-making, and explaining their ideas (Krajcik & Shin, 2014). However, in a semester-long project students’ interest keeps varying due to various reasons, and possibly students might end up losing interest. Sustaining students’ interest in a semester-long project becomes a key challenge. The blind emphasis given to exams or tests acts as an impediment to the learning process where students miss the opportunity to develop an interest in the subject material and don’t prepare for ill-structured real-world problems (Chan et al, 2018).

In spring 2023, the ISD course instructor from a premium educational institute in India proposed a real-world problem of designing a training website for the awareness and sensitization of the institute community towards the challenges and needs of persons with
disability (PwD). Rohm et al (2021) suggest that the authentic real-world problems selected in project-based learning give more relevant learning opportunities to the students. The PwD project was highly relevant for ISD course students and involved stakeholders within the immediate social spheres of students in the institute. The real-time usability of a project by the institute community and the opportunity to make an impact at scale could catalyze students’ interest in the course and project. The PwD project was real-world in nature and demanded a high level of interest and time to execute tasks and achieve a certain level of learning. Thus, the interest-driven creator (IDC) theory was chosen as a suitable basis for the project execution in the ISD course.

Interest is an important component of learning and it is conceptualized as an affective state which can arise from situational triggers (Ainley, 2006). Improved learners’ interest in the topics can make learning easier, more effective, and more enjoyable. According to Chan et al (2018), learning as a process entails interest, creation, and habit, thus learners who develop the habits of creating out of interest will go on to become interest-driven creators. The IDC theory is useful to prepare learners for ill-structured and unpredictable real-world problems (Chan et al, 2018). The interest loop of IDC theory contains three components that are important in sustaining interest. The first component is “triggering interest,” which is facilitating a task that sparks initial interest in learning, the second component is “immersing” learners in learning activities that capture students’ attention and put them in a state of flow, and the third component of the interest loop is “extending” which involve activities to extend learners’ interest refers to integration of prior and new knowledge in learning activities (Wong et al., 2020). The IDC theory further talks about the create loop that includes imitating existing knowledge, combining different concepts to generate new ideas, and staging the work for getting feedback from peers (Chan et al., 2018). The third anchored concept in IDC is the habit loop consisting of a cuing environment to automate behavior, the routine of repetitive behavioral patterns, and harmony as a result of habit activation (Wong et al., 2020). However, the scope of this paper is limited to students’ interest loop hence we will be focusing on how the course-level instructors’ strategies helped in sustaining students’ interest in ISD course.

The research study was set in the premise of the ISD course where research and teaching assistants observed the creation and development of the PwD project (website) as an outcome of students’ interest throughout the semester. The research question (RQ) under investigation is-How did the course level instructors’ strategies help in sustaining students’ interest? To strengthen and validate our observations, we identified the indicators of students’ sustained interest through interviews (with students and instructor), self-report surveys, and in-class observations by researchers.

2. Related Work and Theoretical Basis

To prepare students for the future challenges, the what, why, and how of learning needs to change. An inevitable and permanent change is taking place in learning as it moves from teacher-centered to student-centered learning. Learning becomes enjoyable and effective when students learn with interest, creation further makes learning productive and successful, and instilling the habits of creating with interest can lead to students becoming lifelong interest-driven creators (Chan et al., 2018). Thus, IDC theory focuses on three anchored concepts: interest, creation, and habit. In The Merriam–Webster’s Dictionary “interest” is defined as “a feeling of wanting to learn more about something or to be involved in something.” It has been argued that interest is crucial in guiding behavior and attention and that it promotes learning and inspires effort (Dewey, J. 1913).

The four-phase model for interest development specifies the following phase: triggered situational interest, maintained situational interest, emerging individual interest, and well-developed individual interest (Hidi & Renninger, 2006). Loosely mapped to these phases, the IDC theory proposed a three-component “interest loop” to guide the development of a coherent learning process: triggering interest, immersing interest, and extending interest (Chan et al., 2018). “Curiosity” for triggering interest, “flow” for immersing interest, and “meaningfulness” for extending interest were proposed as suitable indicators by
Wong et al, 2020 with a goal of explaining design factors to be taken into account for early interest development. The action words—triggering, immersing, and extending, call for educators to use them as springboards for student-centered learning that results in deeper immersion in students' interests (Roschelle & Burke, 2019).

IDC theory has been implemented in different regions throughout Asia. In Malaysia, IDC was implemented in educational technology courses for instilling innovativeness among learners (Khambari, 2019) and ignite students' interest in educational technology (Wong et al, 2020). In Taiwan, IDC has been implemented in an experimental school with subjects reading, writing, math, science, English language, and social studies (Wong & Wong, 2019). Kong (2006) used IDC theory to develop computational thinking for K-12 learners. IDC provides a theory of learning design for Asia, hypothesizing that learning activities that are created using the IDC approach will help students become more interested in learning, immerse in the creation process, and reinforce their habit of creating (Chan et al., 2018).

This brings us to the fit of IDC to our own context, where the goal was to create training resources for the institute community to improve sensitization and awareness about the needs and challenges of PwD as a real-world ID project. We have used IDC theory for sustaining students' interest in a semester-long course on instructional system design (ISD).

3. Course Design and Orchestration

Instructional system design (ISD) is a post-graduate level course in the curriculum of the educational technology master and doctoral program. This 6-credit, semester-long course typically covers instructional design (ID) models, tools, and technologies for creating instructional materials, processes followed in ID projects for developing instructional materials, etc. The learning outcome of this course is to enable students to apply given instructional design models. The latest offering of the ISD course was in January 2023 having a total of 22 students registered. Students were categorized into 3 categories- credit, audit and sit through based on their prior experience with the ISD course and registration status. Freshman students attending the course for the first time and to be graded were called credit students. Senior students attending the ISD course for the first time and fulfilling all the course and project requirements except formal grading were called audit students. Students who have attended the course previously, but volunteered to participate in the course and project out of interest were given the role of mentors and called the sit-through students.

The course was delivered in physical mode with a hybrid set-up for smooth execution of the multi-cohort setting. As the course was run in a hybrid setup, the Zoom video conferencing platform was used for synchronous class sessions. Total of 14 classes of 3 hours each were conducted over a period of sixteen weeks. Weekly 3 hours of class time were allotted for both conceptual learning and project work. For the conceptual classes, one classroom was occupied by all students for 3 hours. For the group work activities under the PwD project, 4 different rooms were set up for each group to collaborate with group members. One dedicated research assistant was assigned in each class to record class observations. The physical group work was video recorded and an online meeting was also recorded on the Zoom platform. The recording of data and observations were managed by the research and teaching assistants throughout the course.

Various break-out activities were designed by the instructor for facilitating group discussions during class hours. In each class, 20-30 minutes were reserved for presentations and updates by groups followed by feedback and suggestions by the instructor and students. The instructor has organized three feedback/interaction sessions with subject matter experts (SME), and PwD persons from the institute. The instructor and course team has also organized a 3-hour long workshop by leading organizations in the field of PwD rehabilitation in order to facilitate a deeper understanding of the needs and challenges of PwDs. The class was divided into 4 groups each working on a different type of disability to be covered under the PwD project (website). Figure 1 depicts the different topics covered under each module of the PwD website and the corresponding group composition.
For adding instructional material under each topic, a template that emerged as a result of the class discussion was followed. Each group has used suitable elements for presenting a certain type of content, for example using a poster to depict challenges and support strategies. Combining the inputs and elements proposed by each group, a common template was proposed by the course team for maintaining structural uniformity across the topics. The template divided the website into four sections with the first section covering an introductory video of the topic e.g. “What is attention deficit hyperactivity disorder (ADHD) and learning disability (LD)?”. The second section includes a graphics interchange format (GIF) for depicting the point of view (POV) of a person with a disability. The third section would cover the challenges faced by persons with disability, in the form of a poster, and the fourth section covering support strategies for PwD persons in the institute. The project was situated in an educational institute hence, the key requirement was to contextualize the training resources keeping the institute set-up in mind.

4. Method

4.1 Participants and Data Collection

A total of 22 students registered for the course, of which, 10 students registered for credit, 5 students for audit, and 7 senior students attended the course in the “sit through” mode and contributed towards the PwD project. In the ISD course, stakeholders and end-users were closely connected with students through the PwD website project. Figure 2 gives an overview of the timeline and sources of data collected in the longitudinal study. The research assistants noted in-class observations throughout the course, from the IDC theory implementation perspective. To know the instructors' strategies to sustain interest in the ISD course, we conducted one semi-structured interview with the course instructor at the end of the course.

Figure 1: Types of disabilities covered under the PwD website and group composition

Figure 2: Timeline of course, project execution, and data collection
We have administered an Interest survey to understand students' interest in the course. Students' interest survey questionnaire, designed by a senior post-doctoral researcher who was also a course team member, was administered post-mid-sem exam. This interest survey consists of a total of 7 questions with 4 open-ended and 3 Likert scale questions. For the Likert scale questions, students were asked to respond on a scale of 1-5 corresponding to strongly agree to strongly disagree. We also conducted semi-structured interviews (after the end of the semester) with credit and audit students to understand their interest in the course and PwD project.

4.2 Data Analysis

The PwD project details and subtasks were priorly discussed with all the students in the class. All the ISD students were divided into four teams and each team had 4-6 members. Each team consisted of credit, audit, and sit-through students. For this study, we have sampled credit and audit students. To investigate the students' sustained interest, we have analyzed responses from semi-structured interviews and one survey. The grounded theory approach in content analysis (Mayring, 2015, p. 16) was followed to analyze Interview and survey data. We have done deductive coding of the interview and survey responses. Three components of the interest cycle i.e. triggering, immersing, and extending were used as codes and we have mapped the students’ responses to these codes. Extending the grounded theory approach in content analysis, we have analyzed instructors’ interview responses to investigate the instructor's strategies. To gather additional insights about instructors’ strategy implementation in the classroom and its impact on students’ interest, we have analyzed observational data by researchers following the observational protocol approach (Mayring, 2015, p. 48). The analysis of observational data was exploratory in nature aiming to find more micro-level evidence from the classroom.

5. Findings

This section contains findings on how the course activities and instructors’ strategies help in sustaining students’ interest. Section 5.1 contains an episode showing students’ interaction during a class discussion, mapped to different components of the interest loop. In section 5.2, Table 1 further shows course-level strategies applied by the instructor in the ISD course to sustain the student’s interest in the course. In section 5.3, we show evidence of students’ sustained interest, based on a self-reporting survey questionnaire and student interviews.

5.1 Students’ Interest

The episode below depicts a classroom discussion between the students, where the instructor played the role of a facilitator. This is an example of a carefully designed activity where the instructor triggered students' interest in the task. In this case, it was the creation of training resources for the PwD website. The goal of this discussion was to attain clarity about the task students are about to undertake for the next few weeks. While doing so, the instructor also gave room for students to express their opinions and reach a common understanding for working towards the project goal.

The instructor triggered a discussion by posing the following question prompt: “What is a training resource? Can a brochure or a poster or FAQ be considered a training resource? If yes, under what conditions, and if no, why not? Okay, I don't have an answer. So that's why I'm throwing it open. So that we can collectively come up with an answer to what we are going to call training material. Okay, so this discussion has implications for what we create as material for the PWD cell.”

While putting forward the project requirements, the instructor also reiterated the broader learning goal beyond the project, which is learning ISD concepts through the project.
Above was the anchor question for the discussion and students started participating based on their prior knowledge. The goal of this discussion was to bring out students’ thoughts and views on “What is a training resource?” and make them understand the contextual nature of training resources. Students were seen immersed in the discussion and here are some excerpts from the students’ discourse:

S16: Training through the brochure or poster cannot be considered as a training resource but it depends on what is the title or theme of the poster. If it is a “how to” question then it can become a training resource.

S12: But a poster or brochure can’t be a training module. It comes under awareness. So if it is for spreading awareness about PwD for people other than PwD then it can be considered as a training resource.

S18: I think a poster can be considered a training module depending on what you are trying to learn. For example: “How to use a printer”- a poster is sufficient training for that.

S5: I would like to counter that, yes, the instructions with any equipment or anything can be can be treated as a resource but I can’t be considered training itself. Whatever comes along with the printer or the laptop, it’s a set of instructions or FAQ, It’s not training actually.

S19: I think it depends on the context. For example, during COVID situations, a lot of information was given in the form of instruction through the brochure, and they were kept in public places. Whereas, if the same thing is made in a different context, for example, “You do this so that you will learn how to ride a bicycle” So the instructions given are still instructions but can’t be considered training. So it depends on the context.

Here, the instructor intervened and tried to steer the discussion by extending the point of context. Instructor suggested: “Okay, so let's pause here, for how to ride a bicycle. A poster is not a good training resource. Why? Because the training requires a hands-on and there is no hands-on. That is why it’s not an appropriate training resource. Right. So, that is the point that we have to keep in mind, that it depends upon the purpose. What you have to do is design activities. Okay, so the point of this discussion is that, while we are creating material for the PwD cell, we have to keep asking ourselves this question, what is the purpose of this material? (The implications of this discussion were seen further, during the creation of training resources under the PwD project.)
We present a set of important strategies that were applied by the instructor during the ISD course, along with its goal and evidence that these strategies were useful for sustaining students' interest.

Table 1: course level strategies applied by the instructor to sustain the students' interest

<table>
<thead>
<tr>
<th>Instructors' strategy and a targeted component of interest loop</th>
<th>Goal of the strategy</th>
<th>How this strategy helped in sustaining the interest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Real-world project:</strong> Selecting real-world PwD project with a high social value for the students and institute community. Component: Trigger</td>
<td>To involve the students in a semester-long PwD project</td>
<td>Having a real-world project with implications for stakeholders within the institute, helped students to see the value associated with the project.</td>
</tr>
<tr>
<td><strong>Interactions with experts &amp; stakeholders:</strong> Facilitating Interaction with subject matter experts (SMEs), PwD stakeholders, and experts from the PwD rehabilitation organization Component: Immerse</td>
<td>Students should understand the needs and expectations of stakeholders. Get feedback on work progress, and incorporate the feedback for content refinement.</td>
<td>Students expressed that the real-world nature of the PwD project in the ISD class has played an important role in keeping their interest at a high level. These interactions were perceived as 'great motivations' by the students.</td>
</tr>
<tr>
<td><strong>Group presentations:</strong> Arrange group presentations for peer and instructor feedback post each group activity. Component: Trigger &amp; Immerse</td>
<td>All groups should be aware of the progress with respect to the bigger picture (the whole PwD website) and ideas should be shared to get more clarity</td>
<td>Many valuable suggestions came out of these presentations and discussions. Groups used to discuss challenges in front of their peers used to share potential solutions for the problem. Outcomes from those presentations contributed to the PwD website project. subtasks completion gave them a feeling of confidence and boosted their interest.</td>
</tr>
<tr>
<td><strong>Flexibility in project scope &amp; activities:</strong> Increasing or decreasing the scope of activities dynamically and deciding deadlines with mutual consensus Component: Immerse &amp; Extend</td>
<td>Students should not be overwhelmed by the volume and/or complexity of work, so putting the task within their zone of proximal development to ensure students don’t lose interest.</td>
<td>The PwD project required extra hours of work other than class time. Along with the extra time, it required some content development skills which students had to learn. By involving students in decision-making and discussions, the instructor ensured that the project progress is led and driven by students. This instilled a sense of responsibility and ownership, which helped in sustaining students' interest.</td>
</tr>
</tbody>
</table>
Homework 0:
Asking for students’ reflection after class in and addressing/discussing those points in the next class
Component: Trigger, Immerse & Extend

Giving an opportunity to externalize and put forth the points/opinions. Making students aware of others’ reflections and opinions. Facilitating reflection opportunities to students in asynchronous mode so the learning space remains dynamic.

Students shared that HW0 was like a knowledge-sharing platform, an interesting and encouraging activity they enjoyed. They could read others’ views and were able to comment, hence the reflection activity facilitated the opportunity for the students to know others’ views.

Clear goals:
Mentioning the clear goal and deliverables
Component: Trigger

To bring clarity among students about the nature of the PwD project, deliverables, roles and responsibilities, and time commitments.

Being aware of the goal from the beginning helped students to understand the expected time and effort for PwD project. Having clarity about why and how I should contribute to the course projects and what I will get out of it, helped the students maintain a high level of interest.

Productive digressions: Using digressions as a tool to discuss and connect interrelated knowledge components
Component: Immerse & Extend

To connect other relevant topics with current ongoing topics.

Students expressed that digressions were fruitful for them as it was an opportunity to connect the interrelated topics and construct new knowledge on the basis of prior knowledge.

When asked about the rationale and thought process behind using these strategies, the instructor was quoted saying: “The predominant approach here is to draw upon the knowledge that is present in the class already. So, any orchestration is centered around making people bring out their knowledge and realize that they already know something and also to listen to others and learn from them rather than the approach of the instructor imparting some knowledge.”

While explaining how these strategies were utilized for maintaining students’ interest, the instructor mentioned: “So, maintaining the interest of students is again a fairly challenging task. So, in the beginning, the interest is high and that is automatic, there is some intrinsic interest from the student side. But as the course progresses, a lot of detailing work has to be done. Because of that, it is inevitable that student interest will drop as the course progresses. So, the challenge there was to keep some variety so that the student interest is maintained. So for example, instead of doing all the stakeholder interviews in one shot, we kept looping in that you do something, analyze, reflect, and then do another interview so that each time even though it’s the same project, there is some newness when they are talking to different people at different points of time in the course right. So that was one strategy to maintain interest.”

5.3 Indicators of Students’ Sustained Interest

In this section, we have mapped the evidence of students’ interest in the course to the specific components of the interest loop. This evidence is drawn from students’ responses to the interest questionnaire (administered after mid-sem) and semi-structured interviews (conducted after end-sem). Given below is the evidence of students’ interest viz triggering, immersing, and extending.
Triggering: Three major points were observed to be triggering the initial interest of students in the ISD course. These include, the course content being a valuable addition to students' knowledge as educational technology researchers, the PwD project having real-life value for the institute student community, and instructors teaching methodology. When asked about their motivation for taking the course, students responded:

S17: The motivation for taking this course was to contribute to the project as the project outcome had huge implications on someone's academic life.

S11: The reason for having good interest was the project and the stake of the project because it was very much interested in how it's going to be implemented.

Immersing: For the question: “I enjoy the activities of the class, being unaware of the passage of time”, student responses were seen distributed as strongly agree-7, agree- 4, neutral-4, disagree-0 and, strongly disagree-0. Similarly, for the question, “I find the class activities rewarding in terms of experience”, student responses were distributed as strongly agree-10, agree-5, neutral-0, disagree-0, and strongly disagree-0. Additionally, here are some responses to the open-ended question such as “I look forward to attending every ISD class because ____”. Responses indicate students' immersion in the course:

S4: “In every class, the engagement is different. We will not just sit and look at the slides prepared by the instructor.”

S5: “I will get to learn more about designing various training materials for PwD cell which is aligning with my interest to join this course”.

Extending: For the question: “I find the Pwd project an interesting extension as an application of the learning that is happening in the course”, student responses were seen distributed as strongly agree-11, agree-3, neutral-1, disagree-0, and strongly disagree-0. An example of students extending their understanding is given below:

S13: “PwD project has a non-academic context. We can relate the theories which we are learning, being applied to this project as well. This broadening of the horizon of application of what we are learning beyond the typical academics-related projects gives us a new perspective.”

S11: I had learned during that time that we are not helping them (PwD), it's not like giving them something that they can't do, they are able to do it themselves, we are just giving maybe some sort of scaffold. So this change in perspective and mindset also occurred because of that interaction with the students.

6. Discussion and Conclusion

In this study, we investigated how instructor strategies have helped in sustaining students’ interest in the semester-long ISD course. We have identified the course-level strategies implemented by the instructor and mapped the indicators of triggering, immersing, and extending the interest with the help of the strategies. The findings of this study are in line with Wong et al (2020) and suggest that the thoughtful implementation of IDC theory in the semester-long course can trigger students’ interest in the project activities, immerse them by attaining a state of flow and make them extend their learning in a meaningful way. The findings regarding real-world projects are in line with Rohm et al (2021) and show that the real-world context of the PwD website project has played a bigger role in sustaining the interest. Evidence has shown that having a diverse group composition with students from different levels of expertise gives an opportunity for better collaboration and teamwork.

Each team has put in hours of work while brainstorming ideas, curating and creating training resources, presenting work in front of the whole class, and adopting the changes and feedback given by different stakeholders. Feedback at different time points from the stakeholders, end-users, and experts gave an authentic test bed for the product (PwD website) developed by the students. Working on a relevant problem, acquiring required
skills, managing the task deadlines, and delivering desired product deliverables has provided the ISD students with experience in managing a real-world instructional design project. For students, this experience has been valuable from multiple perspectives, including the learning of ISD concepts, expanding the horizons of understanding the challenges of PwD, and their role as individuals contributing towards a shared goal i.e. PwD website.

For an instructor, understanding students’ interest and aligning it with course projects is an important task. Though it requires intensive efforts, it has helped instructor design effective learning strategies and sustain the students’ interest throughout the course. Sustained students’ interest seems to be a key to sculpting interest-driven creators.

Acknowledgment
We would like to thank the anonymous reviewers, and our study participants for their valuable participation and colleagues for their support throughout the ISD course.

References
Supporting science teachers to select quality edtech learning solutions for their context

Shruti JAIN, Sheeja VASUDEVAN*, Madhuri MUVINKURVE & Sahana MURTHY
Educational Technology, Indian Institute of Technology Bombay, India
*isheeja@iitb.ac.in

Abstract: There has been a surge in the development and use of education technology (edtech) learning solutions in the past decade. Amidst the pandemic, teachers were compelled to incorporate edtech into their instructional approaches. While there are several frameworks, standards, or guidelines available for implementing these edtech solutions, few address the problem of selecting good quality edtech solutions. The ones that address the problem of quality edtech solutions fail to cover all criteria that are needed for effective edtech solutions. In addition, they are not easily usable by the teachers for the selection of an edtech solution. For an edtech solution to be effective it is crucial that it aligns with the context of the learner such as their grade range, their subject of interest, language comprehensibility, etc. Also we found that there are no standards that are available specifically for science teachers in the Indian context. In this paper, we present a validated, reliable, and research-based, Edtech Tulna Index designed to support science teachers in India to select quality edtech learning solutions for their classrooms. We conducted a study to understand the challenges that teachers face and to test the usability of the tulna index and its usefulness to mitigate teachers’ challenges pertaining to the selection of edtech. We found that teachers found the tulna index easy to use. This paper contributes to the research of evaluation of edtech solutions by listing the criteria required for science teachers to select the edtech product, while it also provides the practice community with a usable tool that can be used to select effective learning solutions and potentially enhance their teaching-learning.

Keywords: EdTech, TULNA, guideline, evaluation, useability, perception

1. Introduction

The availability and variety of edtech learning solutions are growing exponentially worldwide. Post-Covid the demand for these edtech solutions is shown to be on the rise and so is the influx of learning solutions into the market (Burns, 2021). The variety in the solutions is brought about not only by the numerous subjects offered, the language of instruction, and age groups but also by different settings in which instruction takes place. Under such circumstances, it is challenging for users (students/ parents/ teachers/ institutions) to choose an effective learning solution (Patel et. al., 2021), i.e. one that contributes to a meaningful learning experience. What is required is a set of guidelines that is easily usable by teachers for identifying and selecting relevant and good-quality edtech solutions. In the absence of such standards or guidelines, the selection will depend on individual teachers’ understanding and perspective of ‘What is a good quality learning solution for science?’ For some, it could be good graphics or 3D animation, while others might feel experimentation or activities are more relevant. This leads to variability and lack of a common understanding of what constitutes ‘good quality’. Often users rely on reviews by previous users and market reviews to guide them in the selection process. Such ad hoc decision-making will lead to selection of an ineffective learning solution that might not best suit their requirement.

Globally there exist some standards and frameworks to evaluate the quality of digital learning solutions. However, these cannot be adopted universally as learning is known to be context-dependent (Bransford et. al., 2000). This paper reports the design and development of a validated and reliable index designed to support science teachers in India in selecting quality edtech learning solutions for their classrooms.
of a research-based, easily usable, context-specific, “Edtech Tulna” index to support teachers in the selection of a good quality edtech learning solution for their use in the science classroom. A study was undertaken with science teachers currently teaching grades 6-10 in India, to explore their perceptions of the index’s usefulness and usability.

2. Literature Review

There are several frameworks and standards that provide guidelines and structures to teachers for effective teaching with ICT. Among these ISTE provides guiding principles for the selection of edtech (ISTE, 2000). CENTA standards (CENTA, 2023) are used for training teachers to use the ICT tools to communicate, create, disseminate, store, and manage information and are aligned with National Professional Standards for Teachers under India’s National Education Policy 2020 (NEP, 2020) guidelines. The Central Institute of Educational Technology (CIET, 2023) provides an e-content evaluation tool and an assessment tool for digital media. Similarly, the recently released Pragyata guidelines for adopting ICT and digital education (Pragyata, 2023) support infrastructure assessment, teacher training, cyber safety and privacy, teacher professional development, physical and mental well-being of students, and a few pedagogical strategies for using ICT. Many of these frameworks focus on effective implementation of ICT and those that focus on the quality of edtech learning solutions do not contain detailed guidelines. Hence the onus is on the teacher to apply the criteria to their subject and context.

During the Covid-19 pandemic teachers were compelled to change their instructional approaches to include digital resources or edtech. There are theoretical frameworks, like UTAUT2 (Venkatesh et. al., 2012), that are developed to understand different aspects of the adoption of technology-based products. In Indian classrooms, the integration of edtech is mainly in the hands of the teacher. According to Bharat Survey for Edtech (2023) Report, about 1 in 3 of 9867 school children received teacher-directed materials through various digital mediums. The context of the learner, the teaching style or method of teaching adopted by the teacher for a particular grade range, the subject being taught, and language comprehensibility are important points to be considered while adopting edtech solutions (Soundararaj et.al., 2022). A study for developing a conceptual framework for evaluating web-based learning resources (WBLR) for school education found the influence of culture on the choice of WBLR among students (Hadjerrouit, 2010). Such studies underscore the need for a framework that considers the context in which the learning is taking place.

Few discipline-specific standards for science such as Next Generation Science Standards (NGSS, 2023) and Common Core Science Standards (2010) are available, however, they are mostly developed in the United States and fail to include the context of learners from other countries. To overcome this problem, countries such as Australia and the UK have come up with their own standards to be incorporated into their curriculum. Since teachers have a good understanding of the context of their learners, it is important to equip teachers with a framework that they can adopt for finding good quality edtech learning solutions for their context. Our research aims at providing this to the teachers in the form of an index called Edtech Tulna (2023) that contains detailed guidelines for teachers in India for selecting science edtech learning solutions in their context.

3. Design of Edtech Tulna Index

We created the Edtech Tulna index as a comprehensive and user-friendly tool for systematically identifying high-quality science edtech learning solutions. Our index is informed by (i) literature from learning theories, multimedia learning, Human-Computer interaction, existing edtech evaluation models, and disciplinary teaching-learning practices in the science domain, (ii) the Indian govt. policies and educational standards and (iii) insights from the product landscape in the Indian edtech ecosystem (Figure 1). The Tulna index is a
rubric covering three constructs: Content Quality, Pedagogical Alignment, and Technology and Design (Edtech Tulna, 2023).

Figure 1. Basis of Edtech Tulna Science Evaluation Index.

3.1 Theoretical Basis for Criteria

Seminal papers with theories related to teaching and learning in the science domain and multimedia learning were used to finalize the twelve criteria in the Tulna Index. The most important criterion for evaluating the content quality of any educational material is the accuracy of the content. Also in science, it is important that the content is up-to-date with recent happenings (e.g., pandemic, etc.) and scientific advancements (NGSS Lead States, 2013). The voice-over and captions used in the content have to be comprehensible by the targeted group of learners for ease of learning (Mayer, 2014; Soundararaj et al., 2022). Alignment with national standards like the National Council of Educational Research and Training (NCERT) and coverage of skills recommended by the government guidelines such as NCF (2005) and NEP (2020), in an Indian context, is also crucial. When designing educational content, it’s crucial to reflect the learner’s local context and ensure inclusive representation in terms of race, gender, religion, and socio-economic backgrounds. This must be achieved while avoiding stereotypes, allowing learners to identify with the content (Dore, 2022).

Table 1. Tulna criteria under each construct

<table>
<thead>
<tr>
<th>Construct (no. of criteria)</th>
<th>Criteria Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content Quality (4)</strong></td>
<td>Content accuracy &amp; clarity, Language comprehensibility, Alignment to national standards, Inclusivity</td>
</tr>
<tr>
<td><strong>Pedagogical Alignment (6)</strong></td>
<td>Content in Context, Learner Scaffolding, Cognitive Engagement, Logical chunking and connectedness, Feedback Quality, Teacher Support</td>
</tr>
<tr>
<td><strong>Technology and Design (2)</strong></td>
<td>Interface Design, Universal Design</td>
</tr>
</tbody>
</table>

Pedagogy for science must include authentic and meaningful scenarios from the local surroundings like the learner’s home, neighborhood, community, and/or culture, that reflect the practice of science (NGSS Lead States, 2013). Support and assistance should be provided to bridge the gap between a learner’s current abilities (actual development) and their potential abilities (zone of proximal development) to accomplish tasks more complex than they could do alone (Quintana et al., 2004; Vygotsky & Cole, 1978). To promote engagement the content should be conversational and visually guided (Mayer, 2014; Van Gog, 2014). Breaking the content into small meaningful segments arranged in logical sequence can enhance learning (Mayer & Pilegard, 2014). Good feedback strengthens the students’ capacity to self-regulate their own performance (Nicol & Macfarlane-Dick, 2006). In addition, instructors need support to integrate Edtech learning solutions into their teaching effectively (Banerjee & Murthy, 2012).
Any technology-based solution should have an easy-to-use and intuitive interface (Davis, 1989). Furthermore, the learning solutions should be designed to be accessible to diverse learners in order to provide equal access and equal opportunity to people with diverse abilities (Caldwell et al., 2008). Based on the literature review and national standards, the final criteria included in the Tulna index are summarized in Table 1.

### 3.2 Structure of the Index

Each criterion in the Tulna Index has easy-to-understand yet detailed reviewer guidelines to help teachers score on a 3 pt-scale of Not-at-all Incorporated, Partially Incorporated, and Fully Incorporated. The reviewer guidelines were developed based on literature and an examination of multiple science edtech solutions for grades 6-8 and 9-10 in the Indian edtech landscape. Insights from exploring existing solutions were utilized to create authentic examples for each criterion. Table 2 gives a glimpse of the index using one criterion from each of the three constructs.

**Table 2. Sample criteria and reviewer guidelines from the Tulna Index**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Reviewer Guidelines</th>
</tr>
</thead>
</table>
| **Language Comprehensibility** | *Is the language appropriate for the intended learners?*  
- The vocabulary and accent used is familiar to the learners of the particular grade range.  
- The sentences spoken or those appearing on-screen are short and easy to follow.  
- The introduction of scientific terms has to be done in a grade-appropriate manner.  
  *e.g. Use of s,p,d,f orbitals to teach covalent bonds in grade 10 is not recommended as the orbitals are not introduced until that grade.*  
- There is no use of slangs or informal language [e.g., gonna, wanna, gotcha] that the learners find difficult to follow.                                                                                                                                                                                                                                                                                                                                 |
| **Content in Context** | *Does the content incorporate appropriate real-life context?*  
- The context provided is relevant to the context of the learner wherever required  
- The context provided is sufficient for the learner to master the specific skill  
- The context does not contain scenarios involving violence, combat, gambling or stereotyping  
  *e.g. to explain the concept of magnetism, showing that a magnet attracts iron nails is relevant but not sufficient, whereas showing that a magnet can be used to separate recyclable iron particles from other wastes in a junkyard is a relevant and sufficient context.*  
- The product provides opportunities to see the application of the skill in the lives of the learners  
- There are multiple ways to make the product accessible to learners with diverse needs.  
  - Audio-only or video-only media provides a transcript and/or on screen text/caption.  
  - Text-only media has a voice-over matching with the text.  
  - Provide captions for all prerecorded/live audio (with or without video) content.  
  - Possibility of recording an answer and hearing assessment questions as voice-overs.  
  - Make it easier for users to see and hear content (*e.g., an option to increase font size*)  

For a particular learning solution if the teacher finds that all the indicators mentioned in the reviewer guidelines are met the learning solution “fully incorporates” the criteria. The learning solution “partially incorporates” the criteria, if only some of the indicators are met or if all the indicators are met but are present inconsistently in the learning solution. And if none or very few of the indicators are met then the learning solution “not-at-all incorporates” the criteria.

### 3.3 Validity and Reliability of Tulna Index

The Face and Content Validity of the Tulna index was established through a rigorous process involving expert review. Individual assessments from the experts comprising a
senior faculty, an edtech expert, and SME were incorporated to arrive at the final index. The interrater reliability was established in 2 stages: (i) during the development of the index and, (ii) after completion of the index. During development, the interrater was conducted among the 3 researchers who were involved in the design of the index. Each researcher assessed 2 learning solutions, yielding an average Cohen’s kappa of 0.335 across the 12 criteria. After this, the researchers discussed their rationale for the assigned scores. Based on the discussion, some modifications to the wording were done and a few more explanations and examples were added to remove subjectivity. After revision, the same team evaluated 2 new learning solutions and the average value of the kappa statistic improved to 0.704, which is considered to be a “fair” agreement (Jonsson & Svingby, 2007). In the second stage, two novice evaluators were trained on the index and they along with a researcher evaluated 4 learning solutions individually. The average value of the kappa statistic between the novice evaluators and the researcher was 0.708, replicating the result observed with the research team thereby establishing the reliability of our index.

4. Study Method

To understand the perceived usefulness, usability, and relevance of the Tulna index we conducted a study with science teachers. The research questions for the study were as follows:

1. Is the Tulna index useful for teachers for selecting edtech learning solutions? How can the teachers adopt it for their use?
2. Do teachers find the Tulna index usable?

4.1 Participants

For purposive sampling, we approached science teachers through social media forums of school teachers, publicizing our study as Edtech Tulna- Sensitization Workshop to aid selection of good quality science edtech learning solutions for their classrooms. The teachers that registered for the workshop were from either government or government-aided schools catering to medium to low-income groups in urban and semi-urban areas in India. Most of the teachers had some experience in using edtech while the number of years of experience varied. A few teachers from non-science domains also registered but their data was not considered for the analysis.

4.2 Procedure

Post registration, the teachers attended a 3hr workshop where they were introduced to the Tulna index, and the 12 criteria were explained in detail with appropriate examples. The workshop was made interactive with several interspersed activities to invite participation from the attendees. Figure 2 shows one of the activities from the workshop.

4.3 Data collection and analysis

To answer the RQs of the study, we employed several data collection instruments and analyzed them using mixed methods (Creswell & Poth, 2016). To get insights into teachers’ backgrounds and perceptions of using edtech in the classroom, we asked the teachers to fill out an online questionnaire during registration. The questionnaire included questions like “According to you what features contribute to good quality content in an edtech learning solution?”, etc.

To understand the perceived usefulness of our index, we conducted semi-structured interviews with interested participants. Perceived usefulness is known to be influenced by culture (Straub et. al., 1997) and it is highly unlikely to be captured through surveys like TAM (Davis, 1989). A few of the questions asked in the interview were, “Did your idea about
evaluating the quality of edtech learning solutions change in any way after the workshop?" and “How does Tulna help you overcome the challenges you face while selecting edtech learning solutions?”. The interviews were recorded and transcribed, which were later analyzed using the principles of content analysis (Mayring, 2015).

To gauge the Tulna index's usability, we had teachers complete the System Usability Scale (Brooke, 1996) post-workshop. SUS is a robust and versatile tool that can be applied to a wide range of ‘products’ (Bangor et. al., 2008). The participants answered the SUS survey containing 10 items on a 5-point Likert scale ranging from strongly disagree to strongly agree. The survey includes both positive and negative questions like - “I thought the Tulna index was easy to use.” and “I think that I would need the support of a technical person to be able to use the Tulna index.”.

5. Findings

5.1 Demographics and background of teachers

A total of 27 educators enrolled in the workshop, with teaching experience spanning from 1 to 30 years. The majority of these instructors were involved in teaching science to students in grades 6 to 8, as well as grades 9 to 10. About 74% (20 teachers) had some experience using edtech learning solutions, such as Audio-Video content, assessments, simulations, and experiments. All the teachers agreed that using Edtech learning solutions was beneficial for the students. When asked to articulate the benefits of using edtech in the classroom, the majority of teachers mentioned that edtech helps in visualization and increases the efficiency of teaching-learning.

Based on responses provided regarding the challenges in selecting edtech learning solutions it was observed that teachers encounter difficulty in finding content for their educational context. A few responses were “Language that is used by majority of the videos is quite not in sync with the demographics that we support”; “The explanation style or presentation style in the available AV are different from my style, that the children are used to for several years.”. In addition, the complexity of selecting a single learning resource amidst an array of options was a concern. The teachers mentioned “I can't choose the source for a particular topic”, and “Selection is not easy as it is not between A or B. Youtube has a lot of videos on the same topic, several channels, etc.”. Spontaneous edtech selection was evident from responses such as “it is time-consuming to check all the content, maybe the first 5 minutes is best but later it is difficult.” and “I select by searching with Google or collecting from friends”.

Figure 2. Sample activity from the workshop on choosing an edtech learning solution
Table 3. Frequency analysis of quality indicators of edtech (reporting indicators mentioned by at least 4 teachers)

<table>
<thead>
<tr>
<th>Quality indicators</th>
<th>freq.</th>
<th>Quality indicators</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>assessment questions</td>
<td>14</td>
<td>integration with other tools</td>
<td>7</td>
</tr>
<tr>
<td>activity and experimentation based</td>
<td>10</td>
<td>gamification</td>
<td>6</td>
</tr>
<tr>
<td>daily-life context</td>
<td>10</td>
<td>user friendly</td>
<td>6</td>
</tr>
<tr>
<td>visualization</td>
<td>9</td>
<td>interesting</td>
<td>4</td>
</tr>
<tr>
<td>easy to understand</td>
<td>8</td>
<td>scalable</td>
<td>4</td>
</tr>
<tr>
<td>build on previous learning</td>
<td>7</td>
<td>curriculum</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3 shows the frequency analysis of the teachers’ responses to “According to you, what features contribute to a good quality edtech learning solution?”. Given that the teachers provided textual answers, their responses were categorized into distinct “quality indicators” in edtech solutions. It is noteworthy that almost no teacher (n<4) mentioned indicators like supporting learning materials, teacher support, short videos, and up-to-date scientific information.

5.2 Usefulness

To gain insight into the usefulness of our Tulna index we conducted semi-structured interviews with 8 out of 16 teachers who attended the workshop. In response to whether the Tulna index aids in selecting appropriate edtech learning solutions, 5 teachers strongly agreed, and 3 agreed on a 5-point Likert scale. One of the teachers said that “Yes, using Tulna criteria gives me faith that I will select the right solutions because when I am reviewing to choose a product it is my perspective or maybe 1-2 teachers but with Tulna tool is made address to a particular solution.” Additionally, teachers were asked about the significance of each of the 12 criteria in selecting edtech learning solutions. Their response was recorded using a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. The outcomes are consolidated in Table 4.

One of the questions that were asked to teachers was “How would you use the Tulna index?”, we found the responses to be very insightful. One of the teachers said that “by understanding criteria such as logical chunking and connectedness, content in context and alignment to national standards, I found the process to make my teaching effective” (translated from Hindi). We also observed a change in the teachers’ perception regarding the quality of edtech learning solutions. When asked the question “After the workshop, has your idea about the quality of edtech learning solutions changed in any way?”, most agreed, and four of the teachers elaborated. One of the teachers said “I can choose the content for my demographics that fits Tulna criteria, focused attention to the particular parameters, and be more mindful while I do my own review.” another teacher said “Yes, I will remember to check for the criterias discussed. Also, I will use the same criteria for video creation. Content accuracy and clarity and proper language are main for any content, logical connections and length of the content to be only 5 to 10 mins for discussing a single concept is very useful. It is useful. Yes, Tulna is useful”. One teacher said “If I follow Tulna criteria then we will get a good outcome… can expect better responses from students. Pedagogy changes with grade but children might get tired of my teaching style so videos will be useful to use in teaching”, while another teacher said “Content if not related to what is not being taught can harmful for children…in point of view of examination the content has to be aligned to national standards. The examples should be given in the context of the child’s learning environment… support should be given in the context of child’s mental ability, long videos should be avoided…”

When asked if “any criteria that you did not check before. But now, after the workshop, you realize that there’s something more that you should look for?”, one of the teachers mentioned that “I will check for content accuracy throughout the entire duration, earlier we used to look for initial few mins… full video must be watched”. Another teacher
said that "There's a difference between the content and the context. So how I can, you know, this is what will be content in the context, making them before making any lesson plan."

Table 4. Avg. score of relevance for each criterion on 5-pt Likert Scale (n=8).

<table>
<thead>
<tr>
<th>Construct</th>
<th>Criteria</th>
<th>Avg. score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Quality</td>
<td>Content accuracy &amp; clarity</td>
<td>4.625</td>
</tr>
<tr>
<td></td>
<td>Language comprehensibility</td>
<td>4.875</td>
</tr>
<tr>
<td></td>
<td>Alignment to national standards</td>
<td>4.625</td>
</tr>
<tr>
<td></td>
<td>Inclusivity</td>
<td>4.125</td>
</tr>
<tr>
<td>Pedagogical Alignment</td>
<td>Content in Context</td>
<td>4.750</td>
</tr>
<tr>
<td></td>
<td>Learner Scaffolding</td>
<td>4.500</td>
</tr>
<tr>
<td></td>
<td>Cognitive Engagement</td>
<td>4.250</td>
</tr>
<tr>
<td></td>
<td>Logical chunking and connectedness</td>
<td>4.500</td>
</tr>
<tr>
<td></td>
<td>Feedback Quality</td>
<td>4.500</td>
</tr>
<tr>
<td></td>
<td>Teacher Support</td>
<td>4.625</td>
</tr>
<tr>
<td>Technology and Design</td>
<td>Interface Design</td>
<td>4.625</td>
</tr>
<tr>
<td></td>
<td>Universal Design</td>
<td>4.250</td>
</tr>
</tbody>
</table>

5.2 Usability

The Tulna index’s usability was confirmed by 8 participant teachers who completed the System Usability Survey during post-workshop interviews. The SUS score of the Tulna index was found to be 71.563, which corresponds to a “good” usability score (Brooke, 1996). The SUS score can be decomposed into usability and learnability (Lewis & Sauro, 2009). The Tulna index scored high on usability (92.969), while it scored poorly on learnability (48.438). Items 4 and 10 measure learnability with statements: “I think I would need the support of a technical person to be able to use this system.” and “I needed to learn a lot of things before I could get going with this system.” Low scores on these two items emphasize the need for Tulna index training.

6. Discussion and Conclusion

To address RQ1, we conclude that teachers found the Tulna index useful, as evidenced by their responses to the Likert scale questions as well as open-ended questions. Teachers stated that they will likely use Tulna in their future practice to identify an edtech product suitable for their needs, thereby making their practice effective. This aligns with prior research which has emphasized the need and importance of contextualization in edtech (Specht, 2006). Teachers also repurposed the Tulna criteria and stated that they will use them as quality criteria during their video creation process. Teachers’ responses were highly reflective and illustrated examples of reflection-on-action (Schon, 1983). We noticed a shift in teachers’ decision-making process from an ad hoc selection approach to a more structured and nuanced one, wherein they applied specific research-based criteria to make a judgment. Teachers also reflected on changes in their own practice, such as viewing content more deeply before making selection decisions. With regard to RQ2, teachers found the Tulna index highly usable, which indicates a high potential for actual use in practice (Pan et al., 2020). Overall, teachers perceived the value of a common set of quality standards for edtech selection rather than varied individual perceptions.

The Covid-19 pandemic has forced educators to adopt and integrate edtech learning solutions into their teaching practices. The lack of appropriate scaffolds complicates the adoption of good quality edtech solutions. Our study introduces a research-based and context-specific Edtech Tulna index, offering teachers a reliable and user-friendly tool to
choose high-quality edtech learning solutions for their classrooms. This study involving science teachers also investigates the shift in their perception of edtech learning solutions following awareness about the Tulna index. Additionally, the study evaluates the Tulna index's usefulness and usability.

Research has established a positive correlation between teachers' perceptions of the beneficial impact of technology and the actual utilization of technology in classrooms (Domingo & Garganté, 2016) but what is missing is context-specific guidelines that would help the teachers to choose relevant quality content for their teaching practices. Available frameworks and standards address the evaluation or implementation of edtech solutions, an index to solve the selection problem of edtech solutions specifically in science education is missing. The Tulna index bridges this gap by providing selection guidelines.

The limitations of our study include the sampling process wherein only the teachers who were already interested in using edtech volunteered and hence our findings could not capture the changes in the perception of teachers who are not inclined to use edtech. Another limitation was that the teachers were mostly from urban and semi-urban schools. It would be beneficial to scale up the study and understand the perspective of teachers from rural India.

The science-specific Tulna index with detailed guidelines on what indicators to check in a learning solution is not only a valuable contribution to addressing the selection challenge faced by teachers but will also give a starting point to the researchers to further explore this area of providing teachers with support for edtech adoption.

Acknowledgments

The authors would like to thank Kavya Alse, Meera Pawar, and Leena Bhattacharya for their contribution to designing the Tulna index. We would like to thank the Centre Square Foundation for the financial support for the project.

References

Caldwell, B., Cooper, M., Reid, L. G., Vanderheiden, G., Chisholm, W., Slatin, J., & White, J. (2008). Web content accessibility guidelines (WCAG) 2.0. WWW Consortium (W3C), 290, 1-34.


A Preliminary Study: Exploring Teachers’ Perspectives on the Role of Gathering Information in Supporting Teachers’ Digital Learning Agility

Kamilah ABDULLAH\textsuperscript{a}, Mas Nida MD KHAMBARI\textsuperscript{a*}, Nur Dania MOHD ROSLI\textsuperscript{a}, Su Luan WONG\textsuperscript{a}, Noor Syamilah ZAKARIA\textsuperscript{a}, Siti Raba’ah HAMZAH\textsuperscript{a}, Priscilla MOSES\textsuperscript{b}

\textsuperscript{a}, Universiti Putra Malaysia, Malaysia
\textsuperscript{b}, Universiti Tunku Abdul Rahman, Malaysia

\textit{khamasnida@upm.edu.my}

Abstract: Gathering information is essential to enhancing teachers’ digital learning agility. By gathering data, teachers can stay abreast of the latest digital technologies and tools, acquire new skills, and improve their teaching practices. This preliminary study explores the role of gathering information in supporting teachers’ digital learning agility. The preliminary study employed a qualitative research approach, using semi-structured interviews to collect data from three teachers. The findings reveal that teachers gather information related to student learning and digital technologies to improve their teaching practices and support their students’ digital learning. The study highlights the importance of information gathering in enhancing teachers’ digital learning agility and provides insights for the development of policies and initiatives aimed at facilitating the improvement of teachers’ digital learning agility.

Keywords: teachers, gathering information, digital learning agility, educational technology

1. Introduction

In today’s digital environment, teachers must have digital learning agility to swiftly and effectively acquire new digital tools and technologies. Chrismastianto and Wibawanta (2023) claim that many teachers lack the knowledge and resources to use digital technology to teach and learn. Getting information is crucial to improving teachers’ digital learning skills. Teachers may develop new skills, keep up with modern tools, and enhance their teaching through gathering information (Van Petegem et al., 2021).

The literature highlights organisational learning as a major consequence of digitalisation (Kuusisto, 2017). This means that organisations, including schools, must constantly adopt new digital tools and technology to stay competitive. Teachers are key to this process as they deploy digital technology and classroom aids. Teachers must gather data to keep current on digital technologies and tools and help their businesses adapt to digitalization.

Additionally, teachers must learn about digital technologies and tools and create digital literacy (Hanan et al., 2022). This encompasses digital information access, evaluation, and organisation. By enhancing their digital literacy, teachers may assist pupils master digital skills for the 21st-century workforce. Digital literacy training and tools may help teachers improve their abilities and promote student digital learning.

Thus, this pilot study interviewed teachers who utilised digital technology in unique conditions (such the COVID-19 pandemic and flooding) to learn about digital pedagogy. The results of this basic study might be utilised to launch a more in-depth study to help policymakers and educators enhance teachers’ digital abilities and teaching efficacy.
2. Literature Review

Since digital technology is essential to education, teachers must be proficient in its use. Information collecting in education is well-known, but how it may assist Malaysian educators become digitally agile is not. Insufficient data address the existing study on Malaysian teachers' digital technology information-seeking practices. However, several studies have examined Malaysian educators' information-seeking. Anuar et al. (2019) found that teachers' attitudes, subjective norms, and perceived behavioural control all affected their intention to use digital technology in teaching. Veeramuthu (2023), on the other hand examined Malaysian literature students' IR4.0 learning practices and challenges. The study found that students struggled with digital literacy and resource shortages while shifting to digital learning. This study found that teachers need digital abilities to help pupils learn in the digital age.

Additionally, Zainal and Zainuddin (2020) examined how national Information and Communication Technology (ICT) regulations affect Malaysia's educational sector, focusing on school technology use. Policy measures improved technological integration in educational institutions, according to studies. Poor infrastructure and teacher training have hindered the optimal use of digital technologies in education. As a result, digital technology has become a crucial tool in education, and teachers must be proficient at using it. There is little information on how Malaysian educators may improve their ICT skills. Several studies have examined Malaysian teachers' digital technology information-gathering practices, underlining the relevance of digital skills in education. These studies suggest that teachers must master digital skills to help kids learn in the digital age. Poor infrastructure and teacher training also hinder the proper implementation of digital technology in education, according to the findings.

Clearly, information acquisition will grow in importance. The role of information collection in shaping instructors' digital learning agility is misunderstood. Thus, to fill this knowledge vacuum, this pilot study answers the following research question: **How does gathering information support teachers’ digital learning agility?**

3. Methodology

This study examined how Malaysian primary school teachers collected data on digital teaching methods amid unprecedented COVID-19 and flood occurrences. The researcher initially contacted prospective instructors for the exploratory inquiry using a prepared information document that explained the study's goals and expected outcomes. Teachers who wanted to participate were screened to make sure they met the researcher's criteria, which included (1) having experience teaching during unsettling times (i.e., COVID-19, the flood, etc.), (2) having taught for at least five years to ensure they had ample experience before, during, and after COVID-19, and (3) teaching at a school with digital facilities. Three Selangor primary school teachers were included in the preliminary research.

After ethical permission, data gathering began. Participants have to sign a form for approval and the Google Meet semi-structured interviews averaged 45 minutes. The researcher copied the interview tapes to her laptop to transcribe and preserve them. The transcriptions were returned to instructors to enhance the interviews and assure accuracy (Mero-Jaffe, 2011).

The data was thematically analysed using the six stages proposed by Braun and Clarke (2006) because it is appropriate for the research as it reports experiences, meanings, and the reality of participants (Braun & Clarke, 2006) in order to uncover teachers' real experiences and gather information about digital teaching.

4. Findings and Discussion

Based on the data taken during the interviews, the findings are grouped in line with the categories of information obtained by the teachers. Notably, when questioned about acquiring information, teachers mentioned two types: (1) getting information and data connected to student learning, and (2) gathering more digital teaching-related information.
Based on the biographical backgrounds gathered, Table 1 depicts the age distribution of three teachers. Their educational backgrounds were as follows: two had bachelor's degrees and one had a master's degree. According to teaching periods, all three teachers had an average of 12 years of teaching experience and had taught before, during, and after COVID-19.

<table>
<thead>
<tr>
<th>Age</th>
<th>30-34</th>
<th>35-39</th>
<th>40-44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

4.1 Gathering Information and Data Related to Students’ Learning

Regarding the first variety of information and data pertaining to student learning collected by teachers, one major theme has been derived from the data collected as presented in Table 2.

<table>
<thead>
<tr>
<th>Students’ Participation</th>
<th>Students’ Proficiency Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Identifying the time and day students are most active</td>
<td>• Plan the teacher’s next teaching strategy</td>
</tr>
<tr>
<td>• Engage with students</td>
<td>• Improvise teaching methods</td>
</tr>
</tbody>
</table>

The teachers were motivated to gather information related to digital teaching because it increases students’ participation in online classrooms. One of the types of information that the teachers gather is the student’s attendance and the teacher explains “...by knowing the rate of attendance on certain days or certain times, we’ll know that, maybe on Wednesday noon, there will be more students.” (T1). T1 also states that once they have identified a suitable time and day that the students are most likely to participate, “…that is the time where you share more input, and for other sessions, maybe you can share something that is not really or very important during that (session)...” They also expressed that by gathering information, they would engage with the students better. T2 explains “...Yes. I keep a record (because) I don’t know the children, so I ask whenever I start to do a reading activity and ask questions or ask what’s your name? I got a list, and then I write down, I record it there because I don’t know the children….” This conclusion is consistent with Fullan’s (2002) study, in which data collecting assists teachers in identifying students’ strengths and shortcomings, allowing them to give focused assistance and interventions.

The second motivational factor was the ability to identify students' proficiency levels which can be done by gathering information related to digital teaching. T2 states that their actions of gathering information after they did online activities helped them to know the students’ proficiency level “…from there, I know their level of proficiency, so then (I know) how to improvise or what shall I do, I mean, my enrichment or remedial activities for the children.” T1 also echoes the same sentiment when s/he states “...from there it actually helps me to plan for my next steps to teach the kids.” This conclusion is corroborated by research done by Basuki and Hidayati (2019), who said that data gathering assists teachers in monitoring student development and identifying students who may be struggling and require further support. They also claimed that data collecting enables teachers to assess the efficacy of educational technology tools and resources (Basuki & Hidayati, 2019).

The researcher found the data insufficient as no participants addressed their difficulties in collecting the material, despite its usefulness in understanding teachers' motivations. The literature has highlighted the challenges teachers face, including (1) time constraints to gather and analyse data; (2) a lack of training or expertise; (3) a lack of technology or assessment tools; and (4) concerns about students' privacy, especially sensitive information. While
gathering data on student learning is crucial for good teaching, the researcher emphasises the importance of addressing teachers' challenges.

4.2 Gathering Additional Digital Teaching-Related Information

Regarding the second variety of information gathered by teachers, namely additional digital teaching-related information, the researcher has identified a central theme as presented in Table 3.

Table 3. Factors Influencing the Acquisition of Additional Digital Teaching-Related Information by Teachers.

<table>
<thead>
<tr>
<th>Keeping up with the most recent digital teaching trends</th>
<th>Enhance teachers’ own knowledge and abilities regarding digital teaching tools and information</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Digital technology constantly changing</td>
<td>• Self-satisfaction</td>
</tr>
<tr>
<td>• Many new digital teaching platforms</td>
<td>• Upgrade own digital technology skills</td>
</tr>
</tbody>
</table>

Based on the data extracted from the preliminary interviews, it is evident that there are several factors influencing teachers’ acquisition of additional digital teaching-related information. P1 and P2 both agreed that there are constant changes when it comes to technology that are teaching-related. P1 explains that “It is important to do this (gather information) because it’s (digital technology) constantly changing and there are also, there are always new things that we need to know that we will learn, or we can learn.” P2 in the interview highlights that “…yes of course. Everything is new. digital itself is very new…” According to Perifanou et al. (2021), by keeping up with digital teaching trends, teachers may meet their daily teaching requirements and prepare for long-term digital education progress. Furthermore, new digital teaching platforms also expose teachers to several digital technologies that might help them boost students’ learning (Hadianti & Rohmah, 2021).

The next influencing factor is to enhance teachers’ own knowledge and abilities regarding digital teaching tools and information. P3 strongly asserts that “We learn digital tools for our own self-satisfaction. We want to know how to use it. It’s very motivating when we feel happy. We like to explore and explore.” When further asked whether P3 always keeps up to date with new digital technology, P3 further explains that “(I) always do. When we discuss innovation, devices, and tools, we typically refer to apps. I seek digital technology-related information through the website. Numerous websites Occasionally through TikTok. TikTok isn’t just about people dancing. There is a video tutorial about design—how to take an angle (when photographing)—so they have instructional videos. Therefore, we update our knowledge of all aspects, not just teaching.” It is important that teachers should follow the latest digital teaching trends to increase their digital literacy and abilities, which can boost student learning (Liza & Andriyanti, 2020; Rahmi & Fajruzzalam, 2022). Additionally, teachers who are competent in using technology for lesson preparation, classroom instruction, assessment and feedback, and contact with students and families are more motivated to learn and use digital technologies (Beardsley et al., 2021).

Using the latest digital teaching methods can have pros and cons. Some teachers lack digital literacy abilities to use technology successfully in the classroom (Liza & Andriyanti, 2020) which may induce frustration and insecurity while utilising digital technology, lowering teaching and learning quality (Shen et al., 2017). A lack of school digital technology infrastructure may also hinder digital teaching (Chrismastianto & Wibawanta, 2023). Some teachers may feel left behind in their digital education developments (Tolks et al., 2020). Finally, teachers may become too dependent on digital technology and neglect other important components of education like building relationships with students and delivering individualised feedback (Lin et al., 2017).
4.3 Teachers’ Preferred Method of Gathering Information

Final data from preliminary interviews are participants' favourite ways of acquiring information. All three individuals preferred to get information online. P1 mentions that the information gathered is "mostly online videos and webinars" while P2 explains that s/he gathered information through “…(I use) YouTube, Google…” P3 on the other hand further expands that “…it’s a good thing (to be able to gather information) because it’s all at your fingertips. By phone. Everything is on the phone, through the Internet.” The discussion with all three preliminary participants highlighted the ease with which knowledge can be gotten in this day and age because everything is available online. It is worth noting that teachers can use the Internet to obtain knowledge on digital teaching (König et al., 2020; Southerton & Lee, 2021; Cahyaningsih, 2021). However, it is critical for teachers to have the appropriate digital literacy abilities to browse and utilise the internet for teaching purposes (Hanan et al., 2022), since this can increase the quality of students' teaching and learning experiences (Liza & Andriyanti, 2020).

The preliminary study shows that acquiring information supports teachers' digital learning agility in numerous ways. First, it keeps teachers up to date on the latest digital tools and technology to better their teaching (Tour, 2016). Second, it helps students find and accumulate learning resources (Eraku et al., 2021). Thirdly, it motivates teachers to evaluate their instruction and improve their digital literacy (Tour, 2016). Finally, collecting information helps teachers socialise and build personal learning networks for continuous digital learning assistance (Tour, 2016). Thus, information collecting is essential to teachers' digital learning agility because it allows them to build digital literacy skills and stay current on the latest digital technologies and tools that might improve teaching.

5. Conclusion

Finally, this pilot research showed how information collecting improves teachers’ digital learning agility. The literature highlighted the relevance of digital skills in education and the challenges to efficient digital technology use in education. Methods detailed this study's qualitative research style, which comprised semi-structured interviews with Malaysian primary school teachers.

The findings show that teachers use student learning and digital teaching information to improve their teaching. The study also identified factors that affect teachers' student learning data collection. The survey found that acquiring information is an important part of a teacher's job and may help them become more technologically agile. Online platforms were the favoured way.

This study expanded teachers’ information collecting and digital learning agility literature. This exploratory study may prompt a more thorough investigation that might inform policies and activities to improve teachers' digital dexterity and pedagogical efficacy. Future study should examine teachers' obstacles in obtaining digital instructional data and how to solve them. Future research should also examine how different information-gathering methods might increase instructors' digital learning agility.

Acknowledgements
This study was funded under Fundamental Research Grant Scheme (FRGS) by Ministry of Higher Education Malaysia. We would like to thank all the teachers for their valuable input to this study.

References


The Role of Flexibility in Shaping Teachers’ Digital Learning Agility: A Preliminary Study

Nur Dania MOHD ROSLI\textsuperscript{a}, Kamilah ABDULLAH\textsuperscript{b}, Mas Nida MD. KHAMBARI\textsuperscript{c,}, Su Luan WONG\textsuperscript{d}, Noor Syamilah ZAKARIA\textsuperscript{e,}, Priscilla MOSES\textsuperscript{f} & Siti Raba’ah HAMZAH\textsuperscript{g}

\textsuperscript{a,b,c,d,e,g} Universiti Putra Malaysia, Malaysia
\textsuperscript{f} Universiti Tunku Abdul Rahman, Malaysia
\textsuperscript{khamasnida@upm.edu.my}

Abstract: Digitally agile teachers are capable to deal with new experience flexibly and rapidly by trying new behaviors and making quick adjustment so that learning can take place even they have constraints especially in using technologies. Flexible teachers are capable to adjust their plan on the fly and make changes to suit current demand. This preliminary study explores the role of flexibility in supporting teachers’ digital learning agility. The study employed qualitative approach with three in-service teachers through semi-structured interviews. Findings suggest that flexibility helped teachers to become digitally agile by developing more effective and strategic behaviors to overcome difficulties in online learning. This study emphasizes the value of flexibility in enhancing teachers’ digital learning agility and offers suggestions for personal and professional development.

Keywords: Teachers, flexibility, digital learning agility, online learning

1. Introduction

The term "flexibility" describes individuals who are able to adapt or adjust to changes in circumstances or expectations, enabling them to remain open-minded and responsive to new ideas, situations, or challenges (DeRue et al., 2012; Burke, 2018; Murphy, 2021). Flexibility is a crucial trait for teachers, enabling them to adapt to changing circumstances and expectations. It allows them to remain open-minded and responsive to new ideas, situations, and challenges. Corpus et al. (2022) found that teachers who can utilize ICT and adjust their teaching methods to suit their students' needs improve learning quality. This paper explores the flexibility of Malaysian teachers as a key factor in supporting digital learning agility and enhancing their performance in the digital age.

Learning agility is a crucial aspect of organizational leadership (Lombardo & Eichinger, 2000; DeMeuse et al., 2010; Bedford, 2011; DeMeuse, 2017a; Burke & Smith, 2018), but it has received less attention in educational settings. Warkentien's (2016) study highlights teachers as strategic classroom leaders, emphasizing teachers are also capable of serving as leaders and can benefit from developing their agility. The importance of flexibility and speed in learning. DeRue et al. (2012) defined learning agility with a primary focus on flexibility and speed. Regarding flexibility, they emphasized the crucial nature of being able to change one's viewpoint to better comprehend how information is related and how it is used to describe a given circumstance. Technology ability is a challenge for teachers in Malaysia, especially during the COVID-19 pandemic. However, a quick response to unexpected situations can help curb learning delays and loss. Thus, exploring teachers' flexibility using qualitative approaches can help develop digital learning agility, enabling them to cope with steep curves and adapt to changing environments.

2. Literature Review
Flexibility is a sign of a better cognitive skills that the capacity to visualize links and move between concepts with ease (DeRue et al., 2012). Flexibility plays a crucial role in enabling teachers to be agile in today’s rapidly changing world, allowing them to grasp evolving educational pedagogies and technologies and adjust accordingly for effective teaching. DeRue et al. (2012) defined learning agility based on cognitive and behavioral processes, with a primary focus on flexibility and speed. In contrast, Burke (2016) presented a different perspective on learning agility, defining it through behavioral components and identifying nine dimensions, including flexibility, speed, experimenting, performance risk-taking, interpersonal risk-taking, collaborating, information gathering, feedback seeking, and reflecting. Despite their divergent approaches, flexibility remains a vital component of learning agility.

Maintaining consistent high performance requires a willingness to learn from mistakes and a positive outlook toward change and failure. These traits underscore the importance of flexibility for agility, as it encourages adaptability in response to situational demands. Allen (2016) examined the value and distinctiveness of learning agility in the workplace and found a significant correlation between flexibility and learning agility. Therefore, it is essential to explore how flexibility shapes teachers’ learning agility, particularly in utilizing digital tools to cope with the ongoing technological advancements in education, to ensure teachers reach their maximum potential in the classroom.

Learning agility is a crucial predictor of leader success (De Meuse, 2017), with only 10% of current leaders having high levels as noted by Joiner and Josephs (2007). This is because, under pressure to complete tasks quickly, people are more likely to make false decisions. Teachers, as classroom leaders responsible for students’ cognitive and social development (Lieberman & Miller, 2005), face the risk of making poor judgments too. Developing learning agility is crucial for motivation, leadership, and engagement (Juhdi et al., 2012). Teachers’ digital learning agility is characterized by speed and flexibility in responding to uncertain circumstances. Strong technology understanding allows them to confidently use various methods (Csordás, 2020; Seale et al., 2010), including modifying, adapting, or personalizing technology to meet current teaching and learning needs. Identifying characteristics that support teachers in becoming digitally agile is crucial.

Research on learning agility in the business sector is extensive, but few studies have explored it in the educational sector. Schools too face dynamic changes due to factors like immigration flows, demographic shifts, and emerging technology (Brzychcy, 2019). The Malaysian Teacher Standard (MTS) 2.0 framework emphasizes four competencies for soulful teachers: knowledge orientation, instructional, community involvement, and personal quality. However, technology expertise is not specifically addressed in this area. To address these gaps and identify the characteristics of digitally agile teachers, the following research question has been posed: How does flexibility shape teachers’ digital learning agility?

3. Methodology

It is worth emphasizing that the actual research focuses on the development of a digital learning agility framework for Malaysian teachers. To meet the study’s aims, a larger set of data was gathered. Nevertheless, a limited scope of the study was extracted for this article, and the objective of this study was to examine teachers’ flexibility in shaping their digital learning agility by examining their experiences with digital tools for online teaching and learning during the school closure caused by the COVID-19 pandemic.

Three Malaysian public-school teachers aged 30 to 44 were selected using purposive sampling, with a minimum of five years of experience or involvement in online teaching during the COVID-19 pandemic. The findings included annotations of the responses from the three teachers, identified as P1, P2, and P3. Semi-structured interviews were conducted to gather data, providing a deeper understanding of the overall experience compared to structured components alone (Moustakas, 1994). End-of-term holiday closures led to interviews conducted via online meeting rooms via Google Meet. Data collection began after ethical approvals from institutions and the Ministry of Education. Participants provided a plain language statement, completed consent forms, and provided audio recordings. Interviews lasted 80 minutes, and audio recordings were transcribed and stored securely.
The interview process involved gathering demographic information, exploring teachers' experiences with technology in the classroom, and their attitude towards being digitally agile teachers. The transcribed data were then structured by using thematic analysis (Braun & Clarke, 2006) as a ‘coding frame’. Data analysis involved recognizing and classifying themes connected to flexibility characteristics that support teachers in becoming digitally agile. Data collected from the participants were coded based on themes and analyzed based on them. Firstly, the phrases and sentences from the data were highlighted and coded according to the meaning of words and sentence structure. The coded data were then consolidated and themes were developed.

4. Findings and Discussion

4.1 Adapting and Thriving in the Digital Classroom

The teachers ranked their digital teaching and learning knowledge as proficient, with most using digital technologies during online learning. They found it helpful, easier, and enhanced their performance. However, their utilization of digital tools decreased after the schools reopened for physical classes. While two teachers continued to employ digital technologies, one teacher has stopped using them. From the thematic analysis, this theme encapsulates the essence of digital agility and flexibility, encompassing the underlying themes of (1) students' engagement, (2) professional growth, and (3) instructional effectiveness. The participants emphasized the necessity for them to adapt their teaching practices to navigate the digital landscape, creating dynamic classroom environments, fostering professional growth, and enhancing instructional effectiveness by adopting an adaptable mindset and leveraging technology for engaging learning experiences.

This section discusses students' engagement, the first theme of the findings. The teachers said that they have used several web-based applications online during class. “While doing the Google Meet, after the session I use Quizziz and usually I’ll ask them to give answers in the chat box.” (P2) and “I used an app on the internet called The Wheel of Fortune. I used the wheel of fortune and made a lucky draw. And I used the medium at the time, Telegram; I used Telegram using video chat, so I used both Telegram and Wheel of Fortune.” (P3). They explained that by trying several platforms, it increases students’ participation and helps teachers' ability to control and manage the class. P2 explains that “… maybe like doing some quiz and then the chat box so they get excited.” P3 in the interview highlights that “When a name was selected, the students got excited. … They’d spread the news, telling their friends that I do lucky draws in my class.” While P1 stated that “…I don’t really know how to switch between devices but after several classes it becomes smoother.” In other words, teachers who are flexible in finding technology integrated strategies in teaching can increase students' engagement. These findings are in line with a study conducted by Park and Park (2020) who found that flexibility helps individual to solve problems, achieve goals and perform effectively.

Professional growth served as the second findings. According to P2, their ability to be adaptable allowed them to gradually overcome unforeseen circumstances “…we can adapt to the situation slowly, but we can adapt. Like myself, so, like we start from zero but everything we learned by ourselves (and) we managed to overcome the situation …”. This findings supported by a study by Barak and Levenberg (2016) which they highlighted that the readiness to participate in unknown settings and enables a learner to efficiently interact with difficult situations. P3 states that online learning requires teachers’ technological proficiency to navigate online tools and thus, by being flexible in responding to this situation, it benefits teachers to expand their digital literacy skills “At that time, most people were more familiar with Google Meet because this feature was very new to Telegram. But after I got the hang of it, it was easier for me.” It is advantageous for an individual's flexibility to successfully cope with online learning to provide them the technological skills essential to handle unforeseen changes (Răducu & Stănculescu, 2021), which supported the finding.

The third findings focuses on the instructional effectiveness which can be done by teachers who are flexible related online teaching and learning. They went through several digital platforms and chose platforms that could be more effective to teach students digitally.
“...try a few apps that is east to be integrated during lessons. So, in the end, I used WhatsApp and then in order to check the answers as well I use another app...” (P1). P1 explained that from the chosen platforms, s/he found it ease students’ evaluation “…so that it’s easier to link the checked worksheet or the products that the kids did to Google Classroom.” P3 stated that online lessons is intimidating but s/he was able to optimally prepared class materials using digital tools “…the preparation was paperless and actually faster.” According to Lübbe et al. (2021), flexibility play a significant role in explaining teachers’ intention to use strategies for students’ achievements and thus, it is evident that flexibility shape teachers’ attitude towards the effective teaching

Despite being able to understand teachers’ motivational factors of being flexible related to teaching online in this study, the researcher noticed that the data inadequate considering only two teachers teach using digital technology after class resumed physically. It is important to note that students currently attending schools that they are characterized as experts in understanding technology (Cruz & Díaz, 2016). Poláková and Klímová (2019) suggest that mobile devices enhance students' knowledge, prompting teachers to update traditional teaching methods and create a modern learning environment. However, challenges in utilizing technology during physical classes should not be overlooked.

### 4.2 Teachers’ Adaptability to Online Teaching

In this study, the teachers expressed various ways of flexibility specially to navigate the digital teaching environment through their adaptability in response to a new and challenging of online lessons. The teachers’ ways of adaptability were divided into two main themes: (1) Teachers as learners, and (2) Supportive network contribution.

All teachers strongly agreed that they had to come out with solutions while teaching online during the pandemic. P1 says that s/he need to find ways to conduct his/her class “…I have to think how to incorporate digital and hardcopy resources…” while P2 says “I have to think something creative to attract the children and focus.” P3 also echoes the same statements when s/he states “…I was looking for a way to get this student to come to my class.” They expressed that through learning, they were able to utilize new digital technologies in online teaching. P1 explains that “I prepared my materials online, but I upload them in the group or Google Classroom … I borrowed my husband’s tablet so that I can use two screens at once to show solutions in doing Math questions … I attend Google Training, webinars, watching videos, that were sent by PPD officers and school.” P2 states that h/she were anxious of using online meeting but managed to conduct class well after several trial and errors “…I did Google Meet first time …I was panicking because it was like trial-and-error session. The, I managed. Actually, I was learning at that time.” It is crucial for teachers to stay current with recent digital teaching needs, by learning, it enables the best adaptability to online learning because the learner and the meaning are prioritized over the content and the activity (Răducu & Stânculescu, 2021).

In addition, the teachers’ adaptability to online teaching were contributed by their supportive network. A teacher commented that his/her colleague help to get knowledges on how to use digital technology to conduct online lesson “Sometimes my friends will give me ideas like why don’t we use this even they said we can play Kahoot games and everything’s right so I learned from friends.” (P2). P3 concurred that using digital technology in partnership with other educators increases their desire in doing so over doing it independently “Like Excel, there are some teachers who I can talk about the internet, the materials they found. When the team is driven, we feel excited, when the team is one-man show, that’s when things become difficult.” It also helps teachers when involves time constraints “…it’s quite helpful because sometimes we don’t really have the time to do everything you don’t have time to even think for that thing so that’s when collaborating actually helps.” (P2). The findings from all three teachers highlighted the ease with which a supportive network truly contributes to the usage of utilizing modern technology throughout the pandemic. It is notable that teachers who work in supportive environments can increase the quality of their teaching in a variety of ways, including through teacher cooperation (Podolsky & Darling-Hammond, 2019).
Adaptability in teachers is important to keep up with constant evolving learning environment especially in mastering technological skills. However, teachers who find this intimidating need support to boost their confident and competence, even post-COVID-19, school might have possibility to close again (Guttentag, 2020). Lieberman (2020) emphasized five strategies to support online adaptability; (1) Providing collegial guidance, (2) Evaluating technology tools collaboratively, (3) collecting data on effective baseline use, (4) Considering online learning applications beyond mandatory, (5) planning intentional technology use to accommodate potential situations.

In summary, based on the findings and discussion of the preliminary study, it is evident that flexibility does support teachers to be digitally learning agile. Flexibility support teachers in many ways. First, being flexible allow teachers to embrace and adopt new technologies by exploring and integrating various digital tools and online platforms (Ulanday et al., 2021). Second, it helps educators to adapt to any changes by staying updated on latest advancements and learn new digital tools (Mardiana, 2020). Third, flexible teachers are able to facilitate collaborative and interactive learning between other teachers or students (Karippanon et al., 2019). Thus, by embracing flexibility, teachers can become digitally agile, adapt to new technologies, exploring digital sources, and customizing instructional delivery in the online teaching environment. This adaptability enables teachers to effectively utilize digital tools and platforms to facilitate meaningful learning.

5. Conclusions

In conclusion, this preliminary study has added to the corpus of knowledge on flexibility and how it affects teachers’ digital learning agility. The literature review has emphasized the value of flexibility in current educational environments and how it might help teachers be nimble. The qualitative research methodology used in this study, which employed semi-structured interviews with public school teachers in Malaysia, has been detailed in the methodology section. The research has shown what motivates teachers to be adaptable when they teach online. The study also revealed how teachers prepared for online instruction and found methods to adjust to unanticipated changes.

Overall, this study has contributed to existing literatures on flexibility and digital learning agility among teachers. The findings of this preliminary study may serve as the basis for the extensive investigation that could provide best practices to be shared with educators, contributing to their professional development and growth. Besides, this study can be beneficial to inform policy makers for policy development and decision-making at various levels. Prior to this study, emphasis was mostly placed on instructors’ experience in supporting their ability to be flexible. Future research is advised to focus on larger concerns, such as the difficulties teachers may have in adapting to the use of digital technology in physical classes and how these difficulties can influence current teaching methods.

Acknowledgements

This study is supported financially by the Fundamental Research Grant Scheme (FRGS) of the Malaysian Ministry of Education (FRGS/1/2022/SSI07/UPM/02/1). The assistance of the Research Management Center (RMC) of Universiti Putra Malaysia in coordinating and distributing funds for this research is greatly appreciated.

References


Incorporation of Robotics in AI Education: Barriers and Enablers as Perceived by K-12 Teachers

Miao YUE\textsuperscript{a}, Morris Siu-Yung JONG\textsuperscript{a,b}, Yun DAI\textsuperscript{a}, Eric Tsun-Hin LUK\textsuperscript{b} & jnalipay@link.cuhk.edu.hk\textsuperscript{b}

\textsuperscript{a}Department of Curriculum and Instruction, The Chinese University of Hong Kong, Hong Kong SAR, China
\textsuperscript{b}Center for Learning Sciences and Technologies, The Chinese University of Hong Kong, Hong Kong SAR, China

*miaoyue@link.cuhk.edu.hk

Abstract: Adopting robotics as a pedagogical tool is commonly observed in school-based programs related to AI education; however, scant research has been conducted on teachers' perceived challenges therein. This study, through interviewing K-12 teachers (n = 11), aimed to probe into their perceived barriers and enablers (as well as the corresponding impacts) in the situation of harnessing robotics in AI education. This study revealed that, apart from the presence of some common external and internal barriers/ enablers, a new factor, namely "technology" concerning the quality and functions of technology emerged. This work provides the field with insights into designing and implementing teacher professional development training on AI education in K-12 contexts.

Keywords: AI Education, Robotics, K-12 Teachers, Barriers, Enablers

1. Introduction

Since the beginning of this decade, artificial intelligence (AI) education has been regarded as a globally important initiative across elementary to higher education (e.g., Chiu et al., 2022; Dai et al., 2023; Li et al., 2022). On top of that, the significant societal impact of AI calls for the incorporation of AI-related knowledge in earlier educational stages. Recently, there has been a notable push to develop AI curricula in K-12 education worldwide, with robotics emerging as a pivotal tool for enhancing learning and teaching (Yue et al., 2022). However, scant attention has been given to the challenges teachers face amid the surge of AI education in the K-12 domain, particularly regarding the use of robotics. While previous research has investigated the barriers related to integrating Information and Communications Technology (ICT) into K-12 education (Brun & Hinostroza, 2014), limited research has focused on the integration of robotics, especially within AI courses. The aim of the present study is to explore the barriers and potential enablers pertaining to robotics-supported AI lessons perceived by K-12 teachers, as well as how these factors might influence teachers' opportunities for teaching with robotics.

AI-powered educational robotics can be an effective tool in teaching K-12 students AI knowledge. Devices such as ClicBot, RoboMaster and LEGO offer various pre-designed modules that support different AI functions, including facial recognition, emotions, sounds, objects, colors, and gesture detection. Teachers can use robotics to increase students' engagement and motivation (Shang et al., 2023; Yang et al., 2023), foster computational thinking (Ching & Hsu, 2023; Shang et al., 2023), and enhance the understanding of programming concepts (Nayak et al., 2023). Despite the above benefits, teachers may encounter various challenges in integrating robotics into teaching practices due to diverse reasons (Ching & Hsu, 2023).
Factors affecting technology integration can be categorized into external and internal. External factors include access to computers or devices, internet connectivity, district policies, teacher training, technical support, peer support, and time. Internal factors encompass personal beliefs, self-efficacy, previous experiences, visions of technology integration, and confidence (Tondeur et al., 2017). Common barriers mentioned include insufficient hardware or software, inadequate technical support, and resistance to technology use. Enablers can include comprehensive professional training, resources, and financial support. There is limited literature that explores the barriers and enablers of integrating robotics into AI lessons.

2. Methodology

Eleven Chinese K-12 teachers (labeled A through K), who are pioneers in school AI education, participated in semi-structured interviews to explore the barriers and enablers of integrating robotics into AI lessons. The interviews were analyzed using a two-step coding approach (Saldaña, 2021). The first phase condensed data into descriptive and inferential codes. The second phase employed pattern and focused coding to generate themes. Constant comparison methods were used to highlight the common barriers and enablers.

3. Preliminary Findings

Two primary external barriers were identified: limited robotics availability and insufficient professional training. As stated by the teacher, “It’s too expensive to provide every student with robotics (B),” the cost of procuring robotics for each student is prohibitively high, impacting their learning experience. Teachers also expressed concerns about the insufficient training to use robotics in their teaching. One remarked, “Without expert guidance, I resort to online resources, which makes me uncertain about the accuracy (M).” A possible reason could be the minor status and positioning of AI as a discipline. AI courses are often embedded within the information and technology (IT) subject, which is viewed as a minor subject. They described IT teachers as, “We are the marginalized teachers who teach the marginalized subject, so school investment in our discipline is severely limited (J).” Eight teachers reported that a lack of confidence in integrating robotics is the primary internal barrier. One remarked, “My academic background is not in AI, and I am not familiar with robotics. Integration of robotics feels risky to me (K).” In fact, it is common for teachers who teach IT, or even AI in K-12, to not have majored in computer science. The lack of AI content knowledge and robotics proficiency contributes to this barrier.

We identified a new category of barrier: the technology factor, which was mentioned by all participants. Different from typical external barriers, such as a lack of hardware (robotics), the technology factor concerns the quality and functions of the robotics. One teacher reported, “Students need one robotic that can carry all AI concepts. I mean, students can learn various aspects of AI through interaction with a single device (D).” However, teachers often need to utilize different kinds of robotics to teach corresponding AI knowledge. Apart from the demand of connecting to different AI content knowledge, teachers expect robotics to create a context for students to solve authentic problems as well. As one commented, “If robotics cannot help me to contextualize a problem to be solved, I will be less inclined to use it in my classrooms (E).” Teachers also emphasized the ability of robotics to capture students’ interests, “The primary aim of AI education for K-12 students is to foster their interests in the AI-powered world. Integrating dull and unintelligent robotics in my classes is challenging because they fail to engage students’ attention on AI technology (F).”

Correspondingly, we also identified enablers under external, internal, and technology factors. Nine teachers acknowledged the effectiveness of robotic competitions as external enablers. One said, “Robotic competitions play a vital role in motivating schools and me to incorporate robotics into the AI learning environment. Students need a platform to examine their learning outcomes, and competitions provide an excellent opportunity to evaluate students’ learning.
performance (C).” Regarding internal enablers, teachers valued positive attitudes towards technology (10 mentions) and constructivist beliefs (9 mentions). One commented, “I firmly believe that students construct their understanding of AI through learning hands-on tasks. By using robotics, students can get an in-depth understanding of the AI problem and deal with these tasks effectively (A).” Finally, all teachers agreed that user-friendly robotics is a critical enabler under the technology factor. They expected robotics to be easy-to-use devices that minimize complications for both teachers and students.

4. Discussion and Conclusion

The preliminary findings highlight key barriers and potential enablers perceived by K-12 teachers when integrating robotics into AI courses, which generally are aligned with previous research on studying external and internal factors of technology integration. Besides, a unique factor, the “technology factors,” emerged from the results of the current study in the context of AI education. Teachers are concerned with the functions and quality of technology (i.e., robotics in this context) itself. The technology factor intersects with both external and internal factors. For instance, teachers believed that if robotics did not possess the expected functions, regardless of the availability of sufficient resources and support (an external factor) or their willingness to use them (an internal factor), it would still present a barrier to teaching AI knowledge. In addition, teachers identified more internal enablers compared to external enablers that facilitate the integration of robotics in AI lessons. The data reveal a strong interest and motivation among teachers towards incorporating robotics in AI curricula, despite any constraints they might experience due to a lack of external support. Future research should aim at designing and implementing teacher professional training programs to address these barriers effectively. Additionally, the impact of the technological factor on teachers' decisions regarding the integration of robotics warrants further exploration.

References

Collaborative design of a simulation-based math classroom: Contradictions and solutions between teaching and research

Wenxin GUO*, Bian WU† & Dong LI‡

†East China Normal University, China
‡East China Normal University, China
¶The School Attached to Shanghai Caoyang No.2 High School, China

* 51214108035@stu.ecnu.edu.cn

Abstract: Although co-design between teacher and researcher as a means could effectively promote teachers’ professional development, little is known about the process of developing a simulation-based math classroom for compulsory education by a co-design team between teachers and researchers. From a dual theoretical perspective involving activity theory and boundary crossing, we investigate the contradictions that emerged between teaching and research activity systems, and the way to deal with the contradictions. The results showed three main contradictions: (1) theoretical ideas versus practical knowledge, (2) simulation-based versus traditional teaching design, and (3) generative interactions versus structured classroom culture. The contradictions were resolved to varying degrees through the efforts of participants who acted as brokers, mainly in terms of the process of perspective making and taking. These findings not only guide how to maximize learning opportunities for teachers in the co-design process with researchers but also enrich teacher education and activity theory.

Keywords: Activity system, contradiction, boundary crossing, math classroom design, mathematics education research

1. Introduction

The increase in digital technologies and teaching tools has placed higher requirements on teachers’ professional capabilities to integrate technology into classroom practice. The collaborative design between teachers and researchers has the potential to improve teachers’ teaching expertise and refine instructional practice by combining the unique expertise of both teachers and researchers (Cai et al., 2017). Due to the different cultural backgrounds of the co-design members, it is common for contradictions to arise in cooperation (Qi et al., 2022). Properly handled contradictions could cause teachers and researchers to critique existing practices or experiences and make an effort to explore new solutions, which in turn can lead to innovation (Yan & Yang, 2019).

In traditional K-12 math classroom teaching, students tend to be polarized in their understanding of mathematical relationships. Some students may struggle with grasping these fundamental principles. Digital classrooms that include interactive simulations are particularly valuable in math teaching and learning (Moeller et al., 2015), especially for internalizing the understanding of complex mathematical concepts and relationships. Currently, research focuses on the results or effects of computer simulations rather than on the enactment process of simulations in actual math teaching during co-design. Therefore, in this study, we aimed to identify the contradictions in the co-design process of a simulation-based math classroom as well as possible solutions to ameliorate them.
2. Literature Review

2.1 Computer Simulations for Math Teaching

Today, technology has become part of students' exploration of knowledge. Computer simulations, one of information and communications technology (ICT) applications, integrate learning areas including cognition, affect, and psychomotility as one and allows students to implement open-ended exploration and systematic experimentation (Ross & Bruce, 2012; Rutten et al., 2015). The specific characteristic of simulation represents complex real-world situations that can be changed by manipulating different parameters (Hillmayr et al., 2020). Through simulations, students could visualize their topics and interact with the technology in doing activities (Garcia, 2020).

Math is most widely known to be challenging for some K-12 students, especially the understanding of abstract concepts and mathematical relationships. Interactive simulation makes those sophisticated math concepts much easier to internalize for almost all students as they observe the direct consequences of the changes they make (Buckley et al., 2004). Most research focuses on the results or effects of computer simulations rather than on the process of teacher enactment of simulations in actual teaching. For example, Rutten et al. (2012) reviewed that numerous researchers have investigated the efficacy of computer simulations without measuring teacher influence or optimal instructional support.

Incorporating interactive digital tools into the math classroom requires the identification of innovative teaching and learning approaches that are suitable for such tools. This process involves identifying course components, making decisions regarding lesson structuring, sequencing, and pacing, as well as instructional strategies for monitoring and responding to student progress (Biggers et al., 2013; Remillard & Heck, 2014). Teachers’ beliefs about the value of classroom teaching and the role of technology in the classroom significantly impact the enactment of these decisions and strategies. In cases that lack alignment, teachers may resist innovations or make substantial adaptations (Bates & Usiskin, 2016; Hermans et al., 2008). The collaborative design between teachers and researchers could be a promising way to facilitate the development of teachers’ adaptive expertise and address theoretical as well as practical issues of teaching and learning (Ko et al., 2022). There are at least two key players who have influence in a teacher-researcher co-design community for the use of computer simulations in math teaching: teachers and educational researchers. In the co-design process, researchers provide educational principles and strategies to facilitate the innovation of teaching approaches and learning techniques. Teachers devise instructional activities to achieve the balance between pedagogical efficiency and innovative teaching practices (Goodyear & Casey, 2015).

2.2 Contradictions and Boundaries in Activity Systems

Teachers and researchers have been guided by their own cognitive paradigms and backgrounds for a long time, each pursuing distinct cognitions, understandings, and objectives in teaching. It is not surprising that certain contradictions may arise during mutual co-design, and the parties involved play a dynamic game in the contradiction (Goodyear & Casey, 2015). Cultural-History Activity Theory (CHAT) views human activity as inherently social and culturally embedded, emphasizing the interplay between individuals and their environment. From the perspective of CHAT, contradictions could be seen as historically accumulating structural tensions within and between activity systems (Ko et al., 2022). Collectives in the same activity system with a common goal orientation communicate, interact, and influence each other. Each activity system is composed of six core elements, including subject, object, community, mediating tools, rules, and division (Engeström, 2001). Based on CHAT framework, Qi et al. (2022) examined the contradictions and their plausible strategies between the mathematics teaching system and university research system, founding some typical contradictions such as tensions between school regulations and the object to promote teachers’ professional development, as well as traditional teaching design versus research-informed teaching design.
The presence of contradictions is not always an obstacle to progress. Instead, they may trigger new attempts and changes by the subjects in the activity systems if identified and resolved properly (Yan & Yang, 2019). Potari et al. (2019) exposed the contradictions across teaching, research, and policy activity systems and their solutions, indicating identification and coordination as two main ways in the teaching and research activity to formulate a collectively meaningful object. In the middle of activity systems, the boundary represents the cultural difference and the potential difficulty of action and interaction across these systems but also represents the potential value of establishing communication and collaboration (Akkerman & Bakker, 2011). When collaborating across the boundaries of activity systems, teachers or researchers may need to enter into unfamiliar territory and combine ingredients from disparate sites to achieve hybrid situations (Engeström et al., 1995). This is commonly referred to as boundary crossing. It could prompt the subjects to question and critique existing practices or experiences, leading them to explore new solutions and ultimately generate innovation.

In our study, the CHAT and boundary crossing would be used to investigate the contradictions that emerge and the way to deal with them during a co-design process of a simulation-based math classroom across two key activity systems: teachers and educational researchers. Two research questions should be addressed:

RQ1: What were the emerging contradictions in the co-design process of a simulation-based math classroom?

RQ2: How did the participants deal with these contradictions between the teaching activity system and the research activity system?

3. Method

3.1 The systemic context and participants

Our co-design team was composed of two groups, including three researchers (one university professor, one teaching and research staff, and one graduate student) and five teachers (three in primary school and two principals). We use the acronyms R (researcher), T (teacher), or their combinations (RT) to refer to the activity systems to which participants belong. All the researchers had extensive research experience in educational technology, which could be seen as full members of the research activity system. These three researchers were represented by the abbreviations R1, R2, and R3, respectively. Teachers could be viewed as full members of the teaching activity system. They were all exposed to periodic teaching research activities with the eagerness to learn about new reform ideas to improve their teaching. Additionally, one of the principals had partial membership in the research activity system as he held a Master’s degree in education and also played a key role in numerous educational research projects focused on classroom teaching supported by interactive simulations. This principal was represented by TR1, and the others were represented by T1, T2, T3, and T4.

The PhET is a site of research-based interactive computer simulation in teaching and learning physics, math, and other sciences (Garcia, 2020). Students engage in exploratory learning through the use of a simulator called 'Number Line: Integers' from PhET, which provides animated, interactive, and game-like environments. The whole team worked collaboratively for about two months (March 2023-April 2023) and prepared three versions of simulation-based mathematics instructional design for trial teaching in three different fifth-grade classes, 35 students in each. After each trial teaching, teachers and researchers attended a seminar meeting to reflect on and improve the current instructional design collaboratively.

3.2 Data collection and analysis

The collaboration took place through three post-lesson meetings (M1, M2, and M3) and numerous meetings in subgroups. After attending the final post-lesson meeting, the teacher
who teaching this mathematics course with the support of PhET was invited to be interviewed. In the following, this interview was represented by the abbreviation I1. The interview questions include four main aspects: the process of using PhET simulation tool in a mathematics classroom; the experience of co-design and improving such a digital classroom; the perceptions of different opinions of teachers and researchers; the meaning of technology innovation and its relation to research or teaching. The data analyzed included the following: (1) video recordings of the three post-lesson meetings, (2) audio recordings of the interview, and (3) three versions of simulation-based mathematics instructional design.

To answer our first research question, the CHAT framework was used to identify and track the emerging contradictions between the teaching activity system and the research activity system in the co-design context. The contradictions were characterized by their content, the involved activity systems, and the elements of Engeström’s (2001) interconnected triangles (see Figure 1).

To address the second research question, the construct of boundary crossing was used to examine how the management of the contradictions by the team members contributed to the design process. Akkerman and Bakker (2011) discerned and summarized four types of learning potential of boundary crossing: (a) Identification involves recognizing boundaries, acknowledging diverse perspectives, and bringing a renewed sense of different practices and related identities. (b) Coordination refers to cooperating efficiently in distributed work and dialogue between diverse partners to maintain the flow of work even in the absence of consensus. (c) Reflection involves recognizing and articulating distinctions between practices, thereby making explicit one’s understanding and knowledge of a particular issue (perspective making), as well as looking at oneself through the perspectives of other worlds (perspective taking). (d) Transformation refers to the engagement of participants from diverse systems in constructive activities that result in significant changes to existing practices and even create novel cultural forms. People who cross boundaries serve as conduits for the introduction of elements from one practice into another, such as brokers and boundary crossers (Star & Griesemer, 1989). The identified contradictions indicated the boundary with the boundary brokers or objects. The process of dealing with this boundary was coded by such four types of learning at the boundaries.

4. Results

We structure the results based on the main contradictions that emerged while preparing and revising simulation-based mathematics instructional design. Three main contradictions were identified and coded, including theoretical ideas versus practical knowledge, simulation-based versus traditional teaching design, and generative interactions versus structured classroom culture. In each of the three subsequent subsections, we present through illustrative examples the identified contradictions, elements involved in activity systems, and how these contradictions were managed through boundary crossing to jointly address research questions one and two.

4.1 Theoretical ideas versus practical knowledge
The first contradiction emerged in the co-design team’s initial post-lesson meeting when they reflected on the implementation of a simulation-based math classroom utilizing PhET. Mathematics teachers made decisions in the teaching process and activities with their stereotypical thinking that relied on the content of mathematics textbooks. At this phase, conflicting perspectives between teaching and research activity systems emerged, indicating the gap between educational theories and practical knowledge. We provide example 1 to illustrate the elements of the activity systems involved in the contradiction and the process for dealing with it.

4.1.1 Example 1

In terms of teaching strategies, R1 suggested incorporating embodied cognition theory in the learning science field into the practical math classroom, and explained how to use, such as “hands-on or body movement”, “changing the size and number of paces to reach a certain number on the number axis.” One of the mathematics teachers T1 questioned the matching of embodied cognition theory with the content sequence of the textbook. T2 who taught this course expressed similar arguments:

Physical activity is not the main focus of this lesson, right? In the first class, students need to know what the number axis is. And then in the second lesson, they should be able to look at numbers on it and describe where they are. Using body language representations would work better for these previous two lessons. But now, students need to compare the magnitude of numbers on the number axis. (T2, I1)

Contradiction: object and tools in research versus tools and rules in teaching
In example 1, it seems that the emerging contradiction is between the object (e.g., using educational theory to transform the mathematics classroom) and tools (e.g., embodied cognition theory) of research activity and the tools (e.g., mathematics textbook) and rules (e.g., organize teaching activities according to the sequence of mathematics textbook) of teacher communities. It appears to be prevalent that such contradiction emerges in the teacher-researcher co-design environment. Teachers often perceive researchers’ solutions as too theoretical and insufficiently practical for implementation in real classrooms (Shrader et al., 2001). In contrast, researchers often view teachers’ limited content knowledge as a hindrance to their effective contribution to co-design efforts (Roschelle et al., 2006). This contradiction reflects the distance between educational theory and pedagogical practice, and the new theory-oriented pedagogy has impacted teachers’ instructional design according to established rules and textbook resources.

Boundary crossing remained in the process of identification
Dealing with the contradiction of “theoretical ideas versus practical knowledge” was a challenge for the co-design team. As a broker at the boundary, the researcher expressed the concept of embodied cognition in a language accessible to teachers and proposed concrete examples to apply this theoretical concept. Teachers were aware of the different perspectives of the researcher and the teacher insisted that the "physical activity" did not apply to teaching the knowledge of comparing the magnitude of numbers on the number axis. The boundary between the two activity systems explicitly signals an identification process demarcating the two communities’ goals, tools, and rules.

4.2 Simulation-based versus traditional teaching design
This contradiction occurred during the first and third post-lesson meetings almost throughout the whole co-design process. The role of educational researchers was critical in formulating the common co-design goals, as researchers identified educational goals and provided PhET tools to achieve them based on their numerous similar research experiences. Teachers have hardly ever experimented with new educational resources in the classroom before, bringing many obstacles to the introduction of simulation resources. We provide two
examples to illustrate divergent views about the balance between simulation-based teaching design and traditional teaching design, as well as the discussions for managing them.

4.2.1 Example 2

The goal proposed by researchers was to use simulation technology to drive innovation in classroom teaching. T2 expressed his confusion about how to achieve it, “We are using PhET software for the first time and have no clue how to use it to help me teach math.” T2 further pointed out that the software was not effective for the first trial teaching, “The software was used less in math classes and did not show its benefits.” After the first post-lesson meeting, TR1 provided T2 with several excellent exemplary lessons integrating PhET resources.

Members of the research group, responding to T2’s concerns, clarified that an inquiry-based teaching approach not only facilitates software functionality but also enhances student motivation:

The teacher does not need to give a concrete hypothesis and ask the students to verify this hypothesis. Instead, students can assume which is greater between a positive number and a negative number by themselves, and get some findings by operating on PhET software. (R1, M2)

T2 mentioned in the interview that advice from researchers and other senior teachers had been instrumental in moving him out of the limitations of traditional instructional design:

After the first trial, several teachers, TR1, and R1 suggested that the students should make full use of the software’s features to explore mathematical rules. I redesigned the inquiry process of students in more detail. For example, I will demonstrate to students how to use this software to explore, and students will learn the teacher’s manipulations to explore the rules of comparing the size of positive and negative numbers and summarize the method of comparing the size. (T2, I1)

4.2.2 Example 3

Teachers, who had higher original expectations for its functionality, were not satisfied with the features currently provided by the simulation resource, such as “the position of the input number cannot be presented on the number axis”. Besides, T2, T3, and T4 expressed their desire for the PhET software to add the function of comparing the magnitude of negative fractions:

Comparing the size of negative fractions is difficult for students. It would be better if the fraction could be represented on this software. Then we can use image examples to help students break through the difficult knowledge point of comparing the size of negative fractions. (T2, M3)

TR1 explained that this is due to the existence of diverse teaching orientations and requirements among different countries, “We started from the Chinese math teaching methods and textbooks, perhaps this software is mainly for game-based learning in other countries”. R1 pointed out that “I will write down the problems we have and try to send an email to the PhET operations team with our suggestions for program improvements.”

In T2’s interview, he reflected on the reciprocal relationship between technology and teaching:

I think that no matter what technology is used, it is ultimately for teaching, and these technologies are designed to help us teach better. The problems we find in the teaching process can sometimes, in turn, drive the development of these technologies. For example, if it can follow the teaching in the improvement, it may become more and more perfect, making it even more suitable for most school teachers to use. (T2, I1)
Contradiction: object and tools in research versus tools in teaching
Interpreting T2’s position, a contradiction could be recognized which is between the object (e.g., technology integration into classroom practice) and tools (e.g., PhET simulation tool) of the research activity system and the tools in the teaching activity system (e.g., the content of mathematics textbooks, traditional teaching approaches). The introduction of advanced technology undoubtedly presented challenges for current teaching practices, as teachers were unfamiliar with its implementation and efficacy (Qi et al., 2022). When teachers feel uncomfortable or unprepared, their ability to instruct with simulations may be diminished (Stern et al., 2008). This contradiction reflects the division between new digital technologies and traditional instructional design. It is not enough for researchers to simply make resources available to teachers. What is more important is to teach teachers how to use these in the classroom to maximize the effect of technology.

Boundary crossing was maintained in perspective making and taking
In dealing with the contradiction, R1 and TR1 act as brokers between research and teaching offering ideas and explanations for teachers to overcome cognitive limitations. R1 introduced the student-centered discovery approach, and TR1 gave teachers excellent examples of lessons that integrate PhET resources, including classroom videos and teaching design. From initial confusion and skepticism about the use of simulation resources in the classroom, the teacher rethought his teaching practices through an educational research lens, incorporating a learner-centered inquiry-based teaching approach and adopting critical perspectives towards technology’s role. This indicates boundary crossing in terms of perspective making and taking. Consistent with existing studies, technology could promote teachers’ pedagogical design capacity and help them design innovative instructional practices to better support student investigations (Beyer & Davis, 2012; Davis et al., 2016). Meanwhile, teachers’ beliefs regarding the role of technology in the classroom are pivotal in shaping their instructional decisions or actions (Bates & Usiskin, 2016; Hermans et al., 2008; Webb et al., 2015).

4.3 Generative interactions versus structured classroom culture
The contradiction between generative interactions and structured classroom culture emerged in the subsequent stages of co-design activities (second post-lesson meeting and third post-lesson meeting). The academic researchers in the team were more focused on generative interactions between teachers and students from the perspective of learning analysis. The emerging contradiction here concerns whether generative discourses were too far from the teaching style and structured classroom culture cultivated over time. The analysis of example 4 illustrates certain elements between teaching and research activity systems involved in contradiction and the process of ameliorating it.

4.3.1 Example 4
R2 pointed out that T2 should capture students’ generative expressions and guide them toward deeper elaboration and expression:

R2 pointed out that T2 should capture students’ generative expressions and guide them toward deeper elaboration and expression:

The resources generated are abundant, but there is one place T2 needs to grab. A student answered that the number is infinite, and T2 couldn’t just sit her down. It should be a little more open in the dialogue between teacher and student, such as a further question that allows students to elaborate and express their ideas in depth.

(R2, M3)

In contrast, T2 responded to the above-mentioned issues in the interview by arguing that these generative conversations were not entirely applicable to teaching his class:

If they don’t know students in my class, their suggestions may not be too in line with the actual situation. I just accept part of it. The students in Ms. Chen’s class are very open and outspoken when answering questions, while the students in my class are
more timid and less outspoken. This is because my usual training requires students to strictly answer questions according to my requirements and rarely think outside the box. (T2, T1)

In the later discussion, T2 reflected that generative interactions have an important role in developing students' divergent thinking. He realized that the current didactic class culture should be changed, "We should make students dare to communicate and think more divergently because students are the center of learning." Teachers should fully anticipate "how students might react in class".

**Contradiction: object and rules in research versus rules in teaching**

In terms of the elements of the CHAT triangle, the goals (e.g., adequate student-teacher generative interaction in the classroom) and rules (e.g., research norms) of the research activity contradict the rules of the teaching activity (e.g., established classroom rules, structured culture). As Reiser et al. (2000) noted, teachers and researchers involved in co-design always follow inconsistent workplace norms which bring challenges. Teachers' focus was on the feasibility of implementation and the matching degree to student characteristics. However, researchers tended to adopt an analytic perspective to focus on goals first. This contradiction reflects the uniqueness of cultural norms in teaching and research, as well as the boundaries between generative interactions and existing classroom cultures.

**Boundary crossing was in the process of perspective making and taking**

In dealing with this contradiction, R2 seems to play the role of boundary people providing practical ideas and strategies (e.g., guiding questions) for T2 to change from structured classroom management forms and teaching paradigms. From the interview data, T2 renewed insight into the importance of generative interactions for developing students' divergent thinking. Also, he demonstrated his willingness to change his classroom culture to make it more divergent and open to generative interactions. This displays the boundary crossing in terms of perspective making and taking.

5. **Discussion and conclusion**

In our study, contradictions emerged in teaching and research activity systems through simulation-based math classroom co-design, and their boundary crossings were explicated. The object, tools, and rules of the research activity appeared to contradict mainly the elements of tools and rules in the teaching activity. The object of the research activity is to transform the classroom with simulation resources to promote inquiry-based learning and understanding of mathematical relationships. The teaching reality and classroom culture posed the researchers with a dilemma in terms of how to communicate clearly with teachers about their implementation of the research-oriented teaching strategies and simulation resources in daily teaching.

Contrary to the results of previous studies, boundary crossing mainly involves constructing a new system that takes different perspectives into account rather than simply questioning (Potari et al., 2019). This is partly due to the strong and rich research base of the researcher group. Teachers trusted the expertise of their co-design colleagues, who acted as brokers to bring plentiful resources (PhET simulation tool, educational theories, principals, strategies, and even examples of course design). When teachers understood and formulated a shared meaningful object with the support of researchers, their developing beliefs about instructional orientation and the value of technology can drive changes in teaching decisions and behaviors that lead to a smooth process of boundary crossing (Webb et al., 2015).

When contradictions emerged, the elements of activity systems (Engeström, 2001) could be used to understand the sources of these contradictions and their boundaries. From the perspective of boundary crossing, we depict the process of overcoming contradictions in the interactions between teaching and research activity systems. Incorporating CHAT and...
boundary crossing, our analysis highlights the dynamic process of simulation-based classroom co-design by culturally diverse groups, in which contradictions inevitably arise. However, the dissolution of contradictions contributes to the establishment of more harmonious collaborative relationships that serve as a driving force for the co-design community. This detailed analysis also depicts how teachers iteratively change their pedagogical conceptions and actions to learn and develop. Based on these findings, we will provide further support for teachers to optimize the learning opportunities offered by the collaborative design and enable them to implement a simulation-based digital math classroom independently.

Acknowledgments
The authors wish to acknowledge the following for their contributions to this research: Zuomin Zhang for help in organizing post-lesson meetings and for participating as a member; Hailong Shen for designing and teaching such simulation-based math course; Liang Li for his feedback about this project and directions for future research; QianYu Zhang for help in organizing the content of three meetings.

Funding
This study was funded by the Special Fund for Postgraduates of East China Normal University to Participate in International Conferences and the sub-project "Empowering Future Teachers with Artificial Intelligence: How to Reconstruct the Role of Teachers" which is part of the National Social Science Fund Major Project "Research on Artificial Intelligence Promoting Future Education Development" (Project No. 19ZDA364) for the year 2019.

References


The Impact of Development-questioning Activities on Students' Pre-writing Ideas

Jia Ling HONG a*, Chang-Yen LIAO b, Tak-Wai Chan a & Jui-Fen Chang a

a Graduate Institute of Network Learning Technology, National Central University, Taiwan
b Department of Hakka Language and Social Sciences, National Central University, Taiwan
*110524025@cc.ncu.edu.tw

Abstract: Studies suggest that questioning is a potent cognitive tool that can guide students toward a more profound engagement with their learning and comprehension, mainly driven by their innate curiosity. In conventional teaching methods, instructors often assign writing topics directly, causing a lack of enthusiasm and hindering students' ability to generate original ideas. This circumstance can gradually breed an aversion to writing. In response to this challenge, our research aims to introduce a question-oriented activity that provides students with a participatory platform for idea generation and organization. By engaging in group collaboration and brainstorming, questioning elements are presented to stimulate students' curiosity. Participation in this activity allows students to compile a broad range of ideas for future writing tasks. The intended outcome is to aid students in developing coherent ideas, bolster their motivation for writing, and refine their skills in questioning.

Keywords: Questioning, Round-robin brainstorming, Pre-writing, Group collaboration

1. Introduction

Hidi's (2006) research emphasizes that curiosity and interest profoundly impact learning motivation. Clark et al. (2019) similarly discovered that student questioning can positively stimulate their curiosity. However, in many Taiwanese classrooms, students avoid asking questions, resulting in an unnerving silence. The reasons behind students' avoidance of questioning could range from fear of inadequate knowledge, potential criticism, or simply a lack of interest in the topic. As prior research has illustrated, questioning skills are vital to continually promote student thinking, facilitate valuable feedback, and enhance topic comprehension, enabling students to uncover more profound knowledge (Wardani, 2014).

Writing is divided into three stages: pre-writing, writing, and post-writing. The pre-writing phase holds significant importance for students as it not only prompts them to explore the topic but also assists them in envisioning how to apply it in their subsequent writing, thus aiding in constructing their content (Hagtvedt et al., 2019). However, in the current education scenario, with an emphasis on the outcomes and quality of writing, the importance of the writing process is often overlooked. Hence, urging students to contemplate the topic before writing can effectively influence their subsequent writing development (Hashempour, Rostampour & Behjat, 2015), sparking curiosity to pose questions and prompting more profound thought about the subject matter.

Given the above, this study centers on pre-writing activities and explores the impact of students' questioning abilities and idea generation before writing. Cultivating a questioning habit can stimulate students' curiosity and enhance their questioning skills, encouraging active writing participation and gradually fostering students' interest in writing (Hidi & Renninger, 2020). With the research question in mind, this study will endeavor to design a Question Generation & Combination (QGC) activity for implementation before writing, offering students an environment to gather and organize their ideas. This activity combines students' curiosity with round-robin brainstorming (Etemadzadeh, Seifi & Far, 2013), guiding students to delve
into and reflect on the topic by continually sparking ideas based on their teammates' thoughts. While curiosity diminishes over time, interest can sustain ongoing exploration (Grossnickle, 2016). Thus, facilitating the establishment of sufficient background knowledge through the activity can smoothen composition in subsequent writing.

2. Literature Review

2.1 Writing Ideas and Challenges

Writing is a form of self-expression, where individuals organize their knowledge to share with readers and develop skills in collecting information. Previous research has divided the writing process into three cognitive stages: planning, translating, and revising (Berninger & Swanson, 1994). It involves planning and generating ideas, organizing thoughts, and transforming them into written expression (Baaijen, Galbraith & Glopper, 2014), followed by evaluating and revising the text. The process of writing instruction typically includes pre-writing, drafting, revising, and sharing (Keen, 2017), with pre-writing being a crucial and often overlooked stage in the writing process. Engaging in pre-writing activities can assist students in planning or organizing their ideas, effectively influencing the development of subsequent writing (Hashempour, Rostampour & Behjat, 2015), and fostering motivation for writing (Mogahed, 2013).

In previous writing experiences, teachers typically adopted a product-oriented approach, focusing on students' writing skills and performance. Students' works were evaluated and commented on based on their content, and students might or might not incorporate the suggestions provided by the teacher. This traditional approach of using prompt-based essays often presented challenges for students, particularly in terms of knowing how to begin their writing. This could be due to students feeling unable to generate ideas, lacking interest in the topic, or not having relevant background experiences associated with the topic. Consequently, students found it difficult to express their creativity within the given time constraints. Over time, due to a lack of ideas and limited curiosity in writing, coupled with insufficient background knowledge, students developed an increasing aversion towards writing.

2.2 Brainstorming

The biggest challenge at the beginning of writing is the lack of ideas, and brainstorming is an effective way to generate new ideas (Hashempour, Rostampour & Behjat, 2015). Scholar Osborn (2012) mentioned that brainstorming enhances creative output, involving four conditions: no criticism of ideas, building on others' suggestions, embracing any unconventional ideas, and generating a large quantity of ideas. By creating a stress-free environment through brainstorming, students can improve their writing skills (Hashempour, Rostampour & Behjat, 2015). Stimulating students' thinking through brainstorming helps them explore ideas related to writing activities and how they can integrate or generate ideas based on their prior background knowledge. Prior to conducting brainstorming, students do not need to make any preparations, and they can freely record their ideas during the process. The purpose is to guide students in utilizing this new thinking approach to generate a multitude of ideas.

In this study, the round-robin brainstorming approach is adopted as one of the methods of brainstorming. According to scholar Roestiyah (2008), it aims to generate ideas within a relatively short period, accumulate ideas, and engage in discussions without being influenced by team members, ensuring that each individual has an equal opportunity to share ideas. Round-robin brainstorming is a method of collaborative idea generation (Mogahed, 2013) that also enhances students' autonomy and sense of responsibility, thereby increasing their motivation (Fawzi & Hussein, 2013).

In the past and up to the present, many students have encountered a lack of inspiration and curiosity towards writing, which limits their imagination. Therefore, this study aims to
stimulate students' curiosity through multimedia materials and generate motivation by prompting them with interesting questions. The round-robin brainstorming method is then employed, giving each student an opportunity to contribute ideas and unleash their creative thinking. By drawing inspiration from the ideas of their team members, who bring different perspectives, the students can explore new directions of thought and collect ideas generated collaboratively. These ideas will serve as the basis for the subsequent writing process in terms of direction and content.

3. Model Process Design

This study focuses on the generation and organization of students' ideas, based on the core concept of the "Interest Creator Theory" (Chan et al., 2018). A "QGC" (Question Generation & Combination) activity is designed and a user-friendly platform is developed for students to use prior to writing. The study adopts a "design research method" with the aim of designing the "QGC" model based on theoretical directions and improving students' writing skills. After establishing the research direction and designing the curriculum, the study is implemented in a primary school teaching environment. The research process, including students' writing progress, is documented and discussed with the research team. After each class, the researchers and the teaching staff discuss the students' learning outcomes and make adjustments to the activities based on the analysis and feedback. The investigation spans one semester and explores topics in social and natural sciences.

The "QGC" (Question Generation & Combination) activity consists of two main components: idea generation and idea organization. It includes four stages: generating ideas, categorizing ideas, integrating ideas, and transforming ideas. By employing the principle of collecting a large number of ideas through posing questions, the aim is to organize scattered ideas, make them more concrete and explicit, and enhance the understanding of the core ideas. After the idea generation stage, group discussions are conducted to categorize and integrate ideas, assisting students in organizing their thoughts. Prior to engaging in the QGC activity, related materials such as multimedia resources can be used. These materials provide inspiration based on different themes, integrating elements like text and images, and conveying information interactively to facilitate student engagement. Through this approach, students can establish preliminary background knowledge and stimulate their curiosity.

In the first stage, idea generation, it is essential to have a plethora of ideas for the subsequent process of organizing ideas. The round-robin brainstorming technique is employed, providing an environment where students take turns writing down their ideas. Through group collaboration, students engage in associative thinking and generate more ideas. The objective is to allow students to generate additional ideas by leveraging the input of others. Additionally, the system platform will offer features for preserving and reviewing ideas, preparing for subsequent collaborative discussions. The following three stages focus on further organizing the generated ideas, namely categorizing, integrating, and transforming ideas. These ideas can stem from the knowledge and experiences of group members. Through group discussions, students organize their perspectives and understandings of the problem, enabling them to consider the problem from different angles. During the categorization stage, all ideas within the group are classified and organized through group discussions. Students can categorize ideas based on their types, sources, and observed characteristics. The system platform assists students in swiftly organizing ideas by dragging and dropping ideas into relevant categories, establishing a conceptual framework of the collected ideas within the group. As students progress through the stages and gain a better grasp of the ideas, they reach the integration stage. Here, they can consolidate seemingly similar ideas, further organizing ideas within each category to reduce redundancy and clutter. The system platform allows students to view the before-and-after merging of ideas, enabling them to freely adjust and create more precise ideas. The final stage, transformation, involves students converting their ideas into open-ended questions. The purpose is to employ questioning techniques to help students gather more information and continuously stimulate their curiosity. Through the collaborative process within the group, ideas can be expanded, leading to the collection of a broader range of ideas. The platform provides transformation
examples and assists in identifying ideas that can be transformed, fostering imagination and creativity during the discovery of new ideas, and developing students' cognitive flexibility.

Through this activity, students may begin to conceptualize the structure of their own writing in their minds and contemplate how to apply the materials to their articles. Despite collecting the same set of questions within the same group, each student's presentation differs. As students organize these idea materials into a coherent essay, they actively structure their thoughts and employ questioning techniques. This encourages students to think more attentively and further promotes their active engagement in the activity.

4. Preliminary assessment

4.1 Research participants

The research participants of this study were fourth-grade and five-grade students from an experimental school in Taoyuan, as well as teachers and researchers. The researchers provide the relevant theoretical basis, and discuss with the teachers the design of the course theme textbooks, the selection of multimedia materials, and the learning status of the students. The selected students are expected to use QGC activities to promote the process of generating their ideas, and through data analysis and physical classroom observation, further adjust the planning of activities suitable for students.

4.2 Data collection and research tools

This study will use quantitative and qualitative analysis, conduct two open questionnaires before and after the activity, including questioning ability questionnaire and writing ability questionnaire, for evaluation and analysis. At the same time, the idea database stored by students in the activity is viewed Source for gathering ideas. In addition, after the event, we will conduct interviews with the students to gain a deeper understanding.

Questionnaire: A test used to assess the quality of students' ability to ask questions about the topic before and after the activity. This questionnaire focuses on assessing students' questioning ideas and the number of ideas, and evaluates the quality indicators of ideas based on three aspects, including topical relevance, fluency and extension.

Writing Ability Questionnaire: It is used to assess students' ability to write descriptions of topic-related pictures within a limited time. This questionnaire will assess the number of words written by students and use punctuation as a unit of calculation to measure the number of ideas in writing.

4.3 Preliminary Results and Discussion

The subject of this study, which carries significant malleability, is the energy component of natural science. We segregate the class into three groups to conduct Question Generation & Combination (QGC) activities. This study tentatively scrutinizes students' questioning ability and its bearing on writing through QGC activities. To decode and analyze the ideas engendered by students, we hark back to the writing creativity indicators and standards employed in earlier studies and define scoring standards for the questions. These include theme relevance, fluency, and originality as three guiding indicators to procure the question quality scores of each group and the entire class for statistical scrutiny, as demonstrated in Table 1.
Table 1. Paired sample T-test results

<table>
<thead>
<tr>
<th></th>
<th>SD</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Average Score</td>
<td>0.64</td>
<td>4.33</td>
<td>30</td>
<td>0.002</td>
</tr>
<tr>
<td>The total amount of question ideas</td>
<td>3.74</td>
<td>3.79</td>
<td>30</td>
<td>0.004</td>
</tr>
<tr>
<td>Write word count</td>
<td>5.21</td>
<td>0.40</td>
<td>30</td>
<td>0.701</td>
</tr>
<tr>
<td>The total amount of Write ideas</td>
<td>5.22</td>
<td>1.88</td>
<td>30</td>
<td>0.093</td>
</tr>
</tbody>
</table>

The principal aim of this study is to probe whether the questioning activities will alter the students' writing and questioning capabilities. By the analytical outcomes of the pre- and post-activity tests, the quality of questions, and the abundance of ideas, the average word count and the number of ideas in the post-test outweigh those of the pre-test. The study identified a significant divergence between the average quality of questions and the abundance of ideas, signifying that students bolstered their questioning skills through the QGC activity (p < 0.05). Even though the influence on writing did not reach statistical significance, the activity facilitated students in reflecting upon the writing process, on average, as indicated by the data.

Most students deem writing their poorest skill, and the shortage of ideas is viewed as one of the vital elements affecting creativity. Furthermore, students are customarily in a state of passive participation in the classroom. Therefore, stimulating thought and soliciting questions in writing activities is paramount. Future research will delve more deeply into the various stages of the movement to fathom how students amplify and organize their ideas during the activity and the linkage between idea transformations and writing. The preliminary research outcomes are elucidated based solely on the results above.

Acknowledgements

This study was funded by the National Science and Technology Council of Taiwan (NSTC 112-2423-H-008-003) and by Research Center for Science and Technology for Learning, National Central University, Taiwan.

References


Prompting Teachers’ Digital Literacy Achievement: A Nationwide Survey of Education Informatization in China

Ziyan CHEa, Jiumei YANGb, Longkai WUC & Di WD

*Faculty of Artificial Intelligence in Education, Central China Normal University, China

Abstract: Education informatization is one of the paramount issues in educational reform. To resonate with Chinese national progress in technology development and national policy for promoting education reform and modernization, this study ascertains the current status of education informatization and teachers’ use of technology in teaching and learning in China. Data from a national survey on primary and secondary schools and teachers about the performance of education informatization was analyzed. According to the results, teachers’ use of technology in teaching is a reinforcement of traditional pedagogy; innovative functions such as the communication and evaluation of teaching and learning are under-utilized. On the basis of this finding, promoting education informatization needs to promote teachers’ coherent Digital Literacy (DL) integration competence. Meanwhile, a professional gap in using digital educational resources was found among teachers. The Ministry of Education, Science & Technology (MEST) and the research team have put sustained efforts to guide the development of education informatization in China, and the results of this study serve as evidence for policymakers and principles to promote teacher professional development in line with the changes of new requirements for teachers’ teaching ability in this information era.

Keywords: Teacher Professional Development, education informatization

1. Introduction

Education has evolved alongside global informatization. In 2019, drawing on a range of educational reforms China has undergone since 1978, the Chinese State Council devised and instituted China’s Education Modernization 2035 Plan (2035 Plan) and the Implementation Plan for Accelerating Education Modernization (2018-2022). These two documents aimed at substantially modernizing China’s education system by 2035. Over the past three decades, the central government strengthened national network coverage, connecting all primary and secondary schools by 2020, with half of them having wireless networks (Dong et al., 2021; Zeng, 2022). Recent focus gradually shifted from device construction to innovating Digital Educational Resources (DERs). To resonate with Chinese technology development and national policy for promoting education reform and modernization, the Ministry of Education has carried out training programmes for teachers and principals in Information and Communications Technology (ICT) in the pedagogical application during the past years (Zeng, 2022).

Even though the nationwide infrastructure has been built, education informatization will be accomplished in actually only when teachers and principals are competent enough to systematically utilize technologies to leverage career-related activities (Kim & Lee, 2011). In pursuit of advancing education informatization, a national survey was conducted by MEST in 2020, focusing on primary and secondary schools. This study presents national survey data concerning teachers’ technology integration in teaching and learning (T&L) and the present state of education informatization in schools. Additionally, this study aims to identify potential areas for improvement and strategic focus to promote education informatization in China based on the current landscape.
2. Theoretical Background

The necessity of education informatization emerges as technology develops. The development trajectory of national policy for education informatization usually follows the sequence from simply using technology and modifying technology to serve the education needs to integrating technology for innovative education practices (Zeng, 2022). Education informatization has had a significant impact on improving classroom teaching. Except for upgrading infrastructure, teachers’ DL is an important step to leverage education informatization (Howard & Tondeur, 2023) for maximum potential.

Empirically, DL encompasses competencies like computer literacy, ICT literacy, information literacy, and media literacy, allowing individuals to access, manage, evaluate, and create information through digital technologies (Law, Woo, de la Torre, & Wong, 2018). Prior research emphasized the importance of effective ICT support in universities for preservice teachers (Wang & Zhao, 2021). Although society and industry introduce new technology to schools, its actual integration is usually reported as a reinforcement of traditional approaches (TAN & PARK, 2016) rather than truly transformative education. Only when the integration and utilization of technology are led by education itself, the education informatization is truly meaningful. True education informatization requires teachers' coherent DL integration competence, fostering productive activities and the creation of pedagogical knowledge (Howard et al., 2023; Kim et al., 2011).

Past research focused on theoretical aspects of China's educational informatization. This study presents up-to-date data and analysis on the current state of China's educational informatization, offering insights to government and higher education decision-makers, aiding policy formulation. This study also adds to the previous research by exploring the mediation and moderation role of teachers' professional background and DERs habits. Three research questions (RQs) underpin this research: RQ1: What is the current DL achievement of teachers? RQ2: What is the current education informatization level of schools? RQ3: What needs to be focused on to advance the DL achievement of teachers?

3. Study Method

Our study was based on the national survey data conducted in 2020 by Chinese MEST as part of a territory-wide evaluation of implementing the Education Informatization Promotion Plan. Considering the large population and enormous education scale in China, for the convenience of analysis, this study selected data from three representative provinces from the eastern (Province A), central (Province B), and western (Province C) regions of China.

3.1 Index Description and Data Analysis

The education informatization index in this study, drawing from prior research (Kim et al., 2011), encompasses three dimensions: accessibility, utilization, and usability. Accessibility refers to the accessibility level of infrastructure (hardware, software, internet connection, multi-media classroom, etc.) established for T&L in schools. Utilization is the extent to which education informatization is employed within schools. Usability is teachers' overall use of technologies, including the use of DERs and software, and the coherent DL integration competence in different T&L sectors, which refers to overall DL achievement in this study.

Diversity of ICT curriculums provided by schools, frequency of the teacher professional development on DL competency, and years of teaching are adopted to explore further possible elements that could advance the DL achievement of teachers.

Data analysis utilized SPSS and Microsoft Excel. Descriptive statistics summarized the questionnaire's primary outcome. Educational background differences were assessed with the Chi-square test. The moderation and mediation effect models were applied to uncover potential moderating and mediating factors affecting teachers' DL achievement.
4. Result

4.1 Descriptive Statistics

A sample size of 131347 in-service primary and secondary teachers completed the online questionnaire, with 103739 (79%) teachers graduating from a teacher education programme. A total of 13352 schools from 57 cities completed the online questionnaire.

4.1.1 Teachers’ DL Performance (RQ1)

Figure 1. indicates most teachers tend to use traditional DERs like Slides in in-class teaching and preparation stage, and use DERs like assessment tool for assessments. The use of DERs in teaching research and experiment are rare.

Figure 2. illustrates the prevalent usage of multimedia productivity tools, professional teaching software, and assessment tools among teachers. The adoption of open educational resources platforms, teaching and management platforms, and online teaching platforms is limited.

Figure 3. indicates an overall DL achievement of teachers that teachers actively utilize information technology in T&L activities, including learning analysis, supportive activities, teaching, teaching optimization, online learning, designing DERs, and teaching sharing sessions. The utilization of technology in other activities is relatively low.

4.1.2 Schools’ Education Informatization (RQ2)
Figure 4. Schools’ Utilization Level of Education Informatization

Figure 4. indicates that at the school level, the utilization of education informatization mainly appears in preparation, teaching, resource sharing, and professional development. Teacher-to-parents and teacher-to-student communication through technology are rare. And the utilization of education informatization in learning analysis and teaching evaluation is low.

4.2 Possible Influence Factors on Teachers’ DL Achievement (RQ3)

The Chi-square test results (Table 1.) indicated a trend toward significance for the effect of educational background on the frequency of using DERs among in-service teachers ($p < 0.001$). Teachers who graduated from the teacher education programme (TEP) had a higher percentage of using DERs more frequently, while teachers who graduated from the non-teacher education programme (NTEP) tended to use DERs less frequently.

Table 1. Group Differences in the Frequency of Using DERs (**p<0.001)

<table>
<thead>
<tr>
<th>Background</th>
<th>Never</th>
<th>Rarely</th>
<th>Some times</th>
<th>Often</th>
<th>Always</th>
<th>Total</th>
<th>Chi-Square test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepar</td>
<td>TEP</td>
<td>2322</td>
<td>5974</td>
<td>10725</td>
<td>34072</td>
<td>50631</td>
<td>103724</td>
</tr>
<tr>
<td>In-class</td>
<td>NTEP</td>
<td>556</td>
<td>1298</td>
<td>2535</td>
<td>8511</td>
<td>14708</td>
<td>27068</td>
</tr>
<tr>
<td>Expere</td>
<td>TEP</td>
<td>1527</td>
<td>5235</td>
<td>10092</td>
<td>30484</td>
<td>56401</td>
<td>103739</td>
</tr>
<tr>
<td>NTEP</td>
<td>345</td>
<td>1213</td>
<td>2451</td>
<td>7788</td>
<td>15811</td>
<td>27608</td>
<td></td>
</tr>
<tr>
<td>Examia</td>
<td>TEP</td>
<td>60907</td>
<td>27556</td>
<td>7975</td>
<td>2482</td>
<td>981</td>
<td>546</td>
</tr>
<tr>
<td>NTEP</td>
<td>15261</td>
<td>7688</td>
<td>2403</td>
<td>797</td>
<td>311</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>Innovat</td>
<td>TEP</td>
<td>30821</td>
<td>30043</td>
<td>17691</td>
<td>8829</td>
<td>4934</td>
<td>3249</td>
</tr>
<tr>
<td>NTEP</td>
<td>7822</td>
<td>7874</td>
<td>4872</td>
<td>2386</td>
<td>1359</td>
<td>926</td>
<td></td>
</tr>
</tbody>
</table>

4.2.1 Mediation Effects on Teachers’ DL Achievements

Table 2 shows the mediating effects of DER diversity on the relationship between software diversity and DL achievements. The direct effects of software diversity on both DER diversity and DL achievements were strong and significant (Coeff. = .590, $p < .001$; Coeff. = .396, $p < .001$). Similarly, the direct effect of DER diversity on DL achievements was also significant (Coeff. = .209, $p < .001$). The mediation condition (Figure 5.) was met as the indirect effect of software diversity on DL achievements through DER diversity was significant.

Figure 5. Hypothesis Model for Software Used, DERs Used, and DL Achievement
4.2.2 Moderating Effect on the Diversity of the Used DERs

Table 3 displays the interaction (Figure 6.) between DL achievement and years of teaching in predicting DER diversity. The interaction effect was significant (p < .05). Simple slope test results indicated a significant positive relationship between DL achievement and DER diversity when years of teaching were low (CI [-0.0031, -0.0016], t = -6.1289, p < 0.05). Years of teaching significantly predicted DER diversity when DL achievement was the same. Less experienced teachers showed higher diversity in using DERs for teaching.

Furthermore, the above mediating and moderating analyses suggest a reciprocal relationship between teachers’ DL achievement and the diversity of used DERs. It proposes that higher diversity of used DERs is indicative of enhancing teachers’ DL achievement, and higher DL achievement leads to increased diversity of used DERs.

![Figure 6. Hypothesis Model for DL Achievement, Years of Teaching and DERs Used](image)

**Table 3. Moderation Effect of Years of Teaching**

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>p</th>
<th>LLCI</th>
<th>ULCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>192.6895</td>
<td>0</td>
<td>8.9412</td>
<td>9.125</td>
</tr>
<tr>
<td>DL achievement</td>
<td>99.4889</td>
<td>0</td>
<td>0.8259</td>
<td>0.8591</td>
</tr>
<tr>
<td>Years of teaching</td>
<td>-5.2969</td>
<td>0</td>
<td>-0.0142</td>
<td>-0.0065</td>
</tr>
<tr>
<td>Int_1</td>
<td>-6.1289</td>
<td>0</td>
<td>-0.0031</td>
<td>-0.0016</td>
</tr>
</tbody>
</table>

4.2.3 Mediation Effects on Teacher Professional Development

Table 4 shows the mediated effects between schools’ utilization level of education informatization and teacher professional development frequency through accessibility level. The utilization level had a strong and significant direct effect on accessibility (Coeff. = 4.311, p < .001) and on teacher professional development frequency (Coeff. = 6.513, p < .001). Similarly, accessibility level had a strong and significant direct effect on teacher professional development frequency (Coeff. = 1.269, p < .001). This meets the mediation condition (Figure 7.) where the indirect effect of schools’ utilization level of education informatization on teacher professional development frequency through accessibility level was significant.

![Figure 7. Hypothesis Model for Software Used, DERs Used, and DL Achievement](image)

**Table 4. Mediation Effect on DL Achievement (via Diversity) (**p < 0.001)**

<table>
<thead>
<tr>
<th></th>
<th>Accessibility level</th>
<th>Teacher professional development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilization level</td>
<td>4.311*** (15.964)</td>
<td>6.513*** (3.823)</td>
</tr>
<tr>
<td>Accessibility level</td>
<td>1.269*** (23.728)</td>
<td></td>
</tr>
</tbody>
</table>
5. Discussion

For RQ1, the descriptive statistics reveal that the most frequently used DERs and software were quite traditional, aligned with previous study (TAN & PARK, 2016). More innovative resources like a micro lecture and teaching and management platform are under-utilized.

Regarding RQ2, the utilization of education informatization at the school level was mainly focused on teaching and teacher professional development. The communication and T&A evaluation functions need to be more utilized.

Regarding RQ3, younger teachers tended to use more DERs, and more diversity of DERs used led to higher DL achievement in different T&L activities, which aligned with the conclusions revealed in previous studies (Wang & Zhao, 2021). This may be due to younger teachers perceived the power of technology in their teacher preparation programs, while older teachers usually received traditional programs. Also, teachers who graduated from TEP showed a higher frequency of using DERs. At school levels, with the same utilization level of education informatization, schools with higher accessibility levels tended to provide more teacher professional development.

6. Conclusion

This paper reports on the current situation of education informatization in China and contributes to revealing important insights for administering education informatization promotion. From a theoretical aspect, observations can be made that teacher professional development is the significant factor that would cause an increase in the use of DERs, leading to higher DL achievements. From the aspect of ICT policy-making, these findings may serve as receptive suggestions for promoting teachers’ DL achievement. The accessibility level of informatization provides chances for teacher professional development, but at the same time, policymakers need to realize that except for infrastructure building, on top of recognition for education informatization, resources should also be allocated for teacher capacity building. After all, the return on infrastructure investment depends on teachers who implement the curriculum.

MEST and the research team have put sustained efforts to guide the education informatization in China and better promote the 2035 Plan. The results of this study serve as evidence for policymakers to focus and reallocate resources for teacher professional development to line with the changes of new requirements for teachers’ teaching ability in this information era.

References

Learning with Conversational AI and Personas: A Systematic Literature Review

Antun DROBNJAK\textsuperscript{a}, Ivica BOTICKI\textsuperscript{a}\textsuperscript{*}, Peter SEOW\textsuperscript{b}, Ken KAHN\textsuperscript{c}

\textsuperscript{a}Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
\textsuperscript{b}National Institute of Education, Nanyang Technological University, Singapore
\textsuperscript{c}University of Oxford, United Kingdom
\textsuperscript{*}ivica.boticki@fer.hr

Abstract: This paper describes the results of a systematic review dealing with the use of personas, avatars, and characters in conjunction with AI-supported tools such as chatbots or generative AI in education. Although the use generative AI in education is gaining traction, this study seeks to systematically review the body of knowledge dealing with personified and conversational approaches to education with both pre-generative and generative AI. The results of the study emphasize the importance of the three key elements of such systems: the use of pedagogical agents, interaction, and personalization. These key elements can be relevant when considering the adoption of the new generation of generative AI in education. Such systems should scaffold learners’ understanding providing guidance and support, promote self-directedness and ensure effectiveness in learning, provide customized learning paths, and promote ethical use.

Keywords: Conversational AI, Generative AI, learning, education, pedagogical agents, personas

1. Introduction

This paper deals with the domain of AI use in education to achieve greater interaction and hence a more immersive educational environment for the students. With the rise of AI in all facets of society, researchers and practitioners have been keen on using it as a tool to supplement pedagogies, which could possibly lead to improved educational outcomes (Huang et al., 2021). The development of AI has had a profound impact on the field of pedagogy by providing new opportunities and opening new challenges. Traditional educational approaches often rely on human teachers delivering instruction, but AI opens up possibilities for automated teaching and content delivery.

With the appearance of generative AI, the potential of AI in education is even more obvious and has gained traction across the globe. In terms of interaction, generative AI has led to the creation of various interactive tools and systems. Additionally, large language models (LLMs), such as the GPT-based models, enable learners to interact with AI systems through natural language, facilitating more nuanced and contextualized conversations.

With generative AI there is a potential to contribute to the improved interaction in the educational process potentially leading to educational benefits (Xu et al., 2022), but pedagogical approaches to achieve so are still work in progress. The challenges become increasingly evident when we delve into the technological peculiarities of generative AI. In this domain, one must closely examine these tools within the context of the machine learning language models they produce, which inherently operate on a probabilistic basis (Lim et al., 2023).
The study is a systematic review study conducted according to the PRISMA methodology yielding a specific set of relevant research papers in the field. The study results indicate that the most relevant studies using virtual agents, avatars, personas, chatbots and other AI-empowered tools in education emphasize pedagogical agents, interaction and personalization as key elements contributing the improved educational results and processes. These are further discussed in the paper.

2. Background

Pedagogy for AI in education refers to the principles and practices that should guide the development and implementation of AI technologies in educational settings. The goal of pedagogy for AI in education is to ensure that these technologies support effective teaching and learning practices and align with the educational objectives (Huang et al., 2021). AI technologies have the potential to revolutionize education, providing students with customized learning experiences and educators with valuable insights into learning processes and outcomes.

According to Järvelä (2006), personalized learning has the potential to enhance student interest and motivation, particularly when students are provided with opportunities to stimulate their creativity and curiosity. Personalized learning can take many different forms, from adaptive learning systems that adjust the difficulty of content based on learner performance to project-based learning that allows learners to explore topics that interest them. For example, according to the findings of Brown et al. (2006), university students expressed a preference for utilizing a personalized virtual learning environment (VLE) for exam revision as compared to non-personalized methods.

Ouyang & Jiao (2021) proposed three paradigms - intelligent tutoring systems, learning analytics, and educational data mining, that can support personalized learning experiences, improve educational outcomes, and inform decision-making in educational settings. These paradigms align with the goals of pedagogy for AI in education, which prioritize the ethical and responsible use of AI technologies while also promoting effective teaching and learning practices.

In their paper Van Brummelen et al. (2021) presented four design recommendations for Human Computer Interaction (HCI), including considering personification, transparency, playfulness, and utility that could improve design of AI-based systems. When designing AI-based learning systems, it is crucial to consider how users interact with the system, what information they need to provide, and how the system provides feedback and guidance. Xu et al. (2022) proposed scaffolding as a solution to improve feedback and guidance. In their paper Crearie (2013) emphasizes the need for HCI to be integrated into the development of such systems and notes that AI-based learning systems have the potential to revolutionize education by providing personalized learning experiences. Main challenge in designing HCI for AI-based learning systems is to balance the need for personalization and standardization.

3. Methodology

The PRISMA method is a systematic approach to conducting reviews that involves three main stages: identification, screening, and eligibility assessment. (Liberati et al., 2009).

The search query was designed to cover the three main areas and to explore the studies at their intersection: personas, characters and avatars; AI-powered tools including chatbots and generative AI tools; and education (Table 1).

Table 1. Search query

<table>
<thead>
<tr>
<th>Search query (Web of Science)</th>
</tr>
</thead>
</table>
The query was applied (in April 2023) to the Web of Science database yielding 576 results (both journal and conference papers). Papers had to focus on the use of AI, chatbots and virtual environments in education and had to be oriented towards the interaction between AI and students. Additionally, the papers should be based on a conducted study rather than being purely theoretical (Table 2).

Table 2. The first set (A) of the study inclusion and exclusion criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 The paper should focus on the use of AI-empowered chatbots, avatars, personas and virtual environments in education</td>
<td></td>
</tr>
<tr>
<td>A2 The paper should include studying interaction between the tools and students</td>
<td></td>
</tr>
<tr>
<td>A3 The paper should be based on a conducted study rather than being purely theoretical</td>
<td></td>
</tr>
</tbody>
</table>

These initially agreed-upon criteria were applied to 80 papers by examining the papers’ titles and abstracts by both coders. The degree of mutual agreement was measured in the form of Cohen’s kappa coefficient (Altman, 1990; Landis & Koch, 1977). Coding resulted in excellent agreement with the interrater reliability for the raters found to be Kappa = 0.91. The selection of the remaining papers was performed by one coder by applying the approach agreed upon by the two coders.

The application of inclusion and exclusion criteria resulted in a set of 56 studies. To extract the most relevant and the most rigorous studies, the second set of inclusion and exclusion criteria was applied to further narrow the results (Table 3).

Table 3. The second set (B) of the study inclusion and exclusion criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B1 Study needs to explore the concept of interactive or personalized learning</td>
<td></td>
</tr>
<tr>
<td>B2 Study needs to measure the impact on learning and wellbeing</td>
<td></td>
</tr>
</tbody>
</table>

This finally resulted in 35 studies for which full-text screening was conducted. The full-text screening was more thorough compared to the title and abstract screening and especially focused on selecting the studies that used techniques based on interaction and personalized learning in improving the educational processes. Following that, 25 papers were excluded, leaving 10 papers in the final list.

The whole PRISMA process is depicted in the diagram in Figure 1.

Figure 1. Prisma diagram flow for the study.
4. Results

A total of 10 studies were selected for close examination as part of this review. All of them focus on education or propose an approach that could be applied in educational scenarios. They are oriented towards all ages of participants including students and children, with sample sizes ranging from 20 to 47. (Table 3).

Table 3. The final set of selected studies (N=10) after the application of the literature review steps

<table>
<thead>
<tr>
<th>Author and year</th>
<th>Topic</th>
<th>Targeted age groups</th>
<th>User study participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terzidou et al., 2016</td>
<td>Pedagogical agents</td>
<td>Students</td>
<td>41</td>
</tr>
<tr>
<td>Pataranutaporn et al., 2021</td>
<td>AI generated characters</td>
<td>All ages</td>
<td>-</td>
</tr>
<tr>
<td>Van Brummelen et al., 2021</td>
<td>Conversational agents</td>
<td>Students</td>
<td>47</td>
</tr>
<tr>
<td>Haller &amp; Rebedea, 2013</td>
<td>Chatbots</td>
<td>All ages</td>
<td>-</td>
</tr>
<tr>
<td>Chien &amp; Yao, 2020</td>
<td>Chatbots</td>
<td>Students</td>
<td>12</td>
</tr>
<tr>
<td>Xu et al., 2022</td>
<td>Interactive video</td>
<td>Children</td>
<td>20</td>
</tr>
<tr>
<td>Park et al., 2022</td>
<td>Humanoid robot</td>
<td>All ages</td>
<td>-</td>
</tr>
<tr>
<td>Almahri et al., 2019</td>
<td>Persona design</td>
<td>All ages</td>
<td>-</td>
</tr>
<tr>
<td>Fu et al., 2022</td>
<td>Knowledge-grounded conversation</td>
<td>All ages</td>
<td>-</td>
</tr>
<tr>
<td>Almahri et al., 2021</td>
<td>Persona design</td>
<td>All ages</td>
<td>-</td>
</tr>
</tbody>
</table>

Overall, the studies problematize interaction between humans and technology as a key to students' better performance and. Xu et al. (2022) suggest that enabling the kind of contingent interaction between child viewers and media characters can bring additional educational benefits not available through standard video programming.

There is an overarching focus on pedagogical agents which are intelligent computer programs that can interact with students to provide guidance and feedback. They are designed to be conversational and engage students in a personalized and interactive learning experience. Terzidou et al. (2016) problematized the effectiveness of pedagogical agents in supporting student learning in online courses and found that the use of pedagogical agents can have a positive impact on students' learning.

Pataranutaporn et al. (2021), Fu et al. (2022), Almahri et al. (2019), Almahri et al. (2021), Park et al. (2022) and Chien & Yao (2020) found that personalization can positively affect students' motivation, engagement, and outcomes while similarly Van Brummelen et al. (2021) found that collaborative virtual environments can enhance student learning outcomes.

Haller & Rebedea (2013) presented a method for building a conversational agent with personality and knowledge about historical figures, which can be used in educational contexts. Similarly, Fu et al. (2022) highlighted importance of agents' personal memory in conversation that can significantly improve their responses. Xu et al. (2022) proposed a more advanced framework for designing pedagogical agents that ask questions, give feedback, use scaffolding, and can adapt to students' learning needs in real-time. The study showed that this kind of framework can improve the effectiveness of pedagogical agents in supporting student learning.

5. Discussion

By integrating pedagogical approaches into generative AI, educators can leverage these systems as valuable teaching tools that support and enhance student learning outcomes. For example, in their study Xu et al. (2022) show that scaffolding, a concept rooted in pedagogy that refers to providing learners with support and guidance as they
progress in their learning journey, can enhance learning experience. In the context of
generative AI, the AI models can be designed to scaffold learners’ understanding by
providing step-by-step explanations, hints, or examples tailored to their individual needs
and can act as virtual tutors offering personalized guidance and support.

Interaction is a key element in the integration of generative AI into educational
settings, enabling learners to actively engage with AI systems in meaningful and
dynamic ways. One form of interactive AI is chatbots, which have gained popularity in
education. Chatbots can serve as virtual tutors or assistants, providing instant
feedback, answering questions, and guiding learners through various educational tasks.
Another form of interactive AI is humanoid robots, which offer a physical presence and
mimic human-like interactions. These robots can engage in conversations, demonstrate
concepts, or facilitate hands-on activities. Language models, such as large language
models (LLMs), are another powerful tool for interactive AI in education. These models
have been trained on vast amounts of data and can process and generate human-like
responses to text inputs. LLMs can engage in natural language conversations with
learners, providing information, explanations, and insights.

Through dialogue and interaction, AI systems can adapt to learners’ responses,
tailor content and activities to their specific needs, and offer customized feedback. The
interactive nature of AI systems fosters active learning, promotes self-directed
exploration, and empowers learners to take an active role in their educational journey.
However, the integration of interactive AI in education should be carefully designed and
accompanied by human guidance to ensure its effectiveness and ethical use.

Personalization is a key objective in education, and generative AI has the
potential to adapt to individual learners, enhancing personalized learning experiences.
By leveraging data and advanced algorithms, generative AI models can analyze student
performance, preferences, and learning patterns. This information can then be used to
tailor the content and delivery of educational materials to meet each student’s specific
needs. The configuration and adaptability of generative AI models allow for dynamic
adjustments, providing customized learning paths, adaptive assessments, and
individualized support. This personalized approach empowers students to learn at their
own pace, explore their interests, and maximize their learning potential.

6. Conclusions

The development of generative AI has brought the potential for significant
advancements in the fields of pedagogy, interaction, and personalization in education.
By integrating pedagogical principles into AI systems, educators can leverage
generative AI as powerful teaching tools, supporting and enhancing student learning
outcomes. The interactive nature of AI systems, such as chatbots, humanoid robots,
and LLMs, promotes engagement and facilitates dynamic learning experiences. These
systems allow learners to actively interact with educational content and receive
immediate feedback fostering active participation and deeper understanding. The
combination of pedagogy, interaction, and personalization in generative AI holds great
potential to revolutionize education, providing learners with more effective, engaging,
and personalized learning experiences. However, it is important to ensure that the
integration of AI in education is accompanied by ethical considerations and human
guidance to maximize its benefits and mitigate any potential challenges. With further
research, development, and collaboration between AI technologies and human
educators, the future of generative AI in education looks promising, with a
transformative impact on teaching and learning.
References


Supporting Content Creators in Creating Accessible Digital Materials in Higher Education

Eirik HANSSENa & Weiqin CHENa,b*

aOslo Metropolitan University, Norway
bUniversity of Bergen, Norway
*weiche@oslomet.no

Abstract: The recent growth of the digitalization of education has increased the importance of digital accessibility. In higher education, faculty members and administrative staff are the main content creators who are responsible for creating and publishing digital content for both students and university staff. Literature shows that these content creators do not have the necessary knowledge to create accessible digital content. At the same time, there is an abundance of information available online about how to create accessible digital content. However, due to the lack of time the content creators are often not able to find the necessary information to help them with the particular accessibility task they have at hand. In this study we have developed an online guide aiming at assisting content creators to create accessible content where target users have been involved in the iterative design, development and evaluation process. A summative evaluation was also conducted with a mixed-method approach, combining observation, survey, semi-structured interview, and document inspection. The results show that the online guide has potential to help faculty members and administrative staff to improve the accessibility of their digital content.

Keywords: Digital accessibility, content creator, higher education, summative evaluation

1. Introduction

The recent growth of the digitalization of education has increased the importance of digital accessibility (Chen, 2022; Lazar, 2022) which refers to the design and development of digital content, technologies, and environments that can be accessed and used by all people, regardless of their disabilities or impairments. Many students who have previously experienced barriers in the digital part of their education risk complete exclusion from the university experience after the education has gone mostly or entirely digital (Lazar, 2022).

Digital technologies and content that students use include websites, e-books, content/learning management systems for courses, registration mechanisms, video and other multimedia content (Lazar, 2021), as well as learning materials and documents in various formats such as .docx and PDF, presentations such as PowerPoint files. In higher education, faculty members and administrative staff are the main content creators who are responsible for creating and publishing digital content for both students and university staff. Content creators use many different languages, platforms, devices, operating systems, and cloud-based as well as locally installed software when creating digital content.

Literature has shown that content creators in higher education, such as faculty members, do not have the necessary knowledge to make their digital content accessible (Sanderson et al., 2022). There is already a wealth of online resources where one can learn about digital accessibility and how to make content accessible such as the information provided by the Web Accessibility Initiative (WAI) and many YouTube videos. However, due to the lack of time, it is often difficult for faculty members and administrative staff to look for the necessary information that can help them with the particular accessibility task they have.
at hand. It is, therefore, necessary to provide customized guidance and support to content creators so that they can gain practical knowledge in the process of making digital materials accessible.

The goal of this research is to better support content creators in creating accessible digital materials in the context of higher education. We developed an online guide for assisting content creators and conducted a summative evaluation with the target users to understand how the online guide can contribute to improving the accessibility of the content created by faculty members and administrative staff in higher education.

2. Related Research

Digital technologies have been viewed in scientific literature as resources for achieving inclusion education, both in virtual and in-class environments (Perera-Rodríguez & Díez, 2019). However, studies have shown that digital technologies create new barriers for students with disabilities. Inaccessible information and technology have been found to be one of the main barriers for students with disabilities in higher education (Moriña, 2017). The inaccessible digital materials, including peer-reviewed journal databases and digital tools and environments, hinder the full participation of students (Seale et al., 2015). Fichten and colleagues (2022) argued that future publishers, web designers and educators must be trained on the fundamentals of accessible documents and web design.

Several studies examining accessibility and inclusion in higher education have highlighted that faculty members and administrative staff lack the necessary knowledge of accessibility (Shinohara et al., 2018). A more recent study (Sanderson et al., 2022) conducted interviews with 35 faculty members in the faculties of computer science and engineering. The findings from the study showed that most participants lack sufficient understanding of digital barriers and practical knowledge on how to make digital learning materials and courses accessible. Providing training to faculty members and administrative staff is considered essential to increase competencies in digital accessibility (Bong & Chen, 2021). However, despite of the willingness to learn, faculty and staff were found in lack of time and resources to gain knowledge about and implement digital accessibility in addition to their already full workload (Sanderson et al., 2022).

3. The Interactive Guide for Content Creators

The Interactive guide is a website where content creators can look for guidance for creating accessible digital materials based on the task they are working on.

The first step in the design of the interactive guide is to gather the types of digital content often used in higher education and the accessibility requirements and guidelines. Digital content comes in many file formats ranging from word processor documents, spreadsheets, and presentations to web-based content presented in text and structured HTML, with presentation styled in CSS and possibly being made interactive either using native constructs of the HTML language or by being supplemented with JavaScript. Multimedia such as bitmap images, Scalable Vector Graphics, audio, and video are presented within different types of files and web pages. In addition to accessibility guidelines and standards such as the Web Content Accessibility Guidelines (WCAG), which have clear requirements concerning digital content, different software, platforms, and operating systems have also developed their own guidelines for creating accessible content. For example, Microsoft has supporting videos and manuals on how to make accessible Word documents and PowerPoint presentations for different versions and operating systems. Apple has provided guidelines for accessible documents and presentations.

The interactive guide was developed through an iterative process. Three iterations of prototyping and user testing were carried out. During the iterations three participants were involved in the testing process and gave feedback on the design, content, and organization of the content in the interactive guide (uuguiden.no). Figure 1 shows that users can apply
filters to the guide to get the necessary information related to the task they are working on. The filters include the types of content and the software they use, as shown in Figure 1. Another option for users is to browse a table of contents.

![Figure 1](image.png)

*Figure 1.* The interface where users can apply filters to the guide.

## 4. Evaluation

### 4.1 Method

The summative evaluation included two parts which collected both qualitative and quantitative data. The first part was a usability study with observations and interviews. The participants were asked to test the interactive guide by completing some tasks in creating accessible digital content or finding information about making a component (e.g. table, image) accessible. Observation notes were taken during the testing, followed by a semi-structured individual interview. The interview questions focused on the participant’s experience with the interactive guide. The interview was audio recorded and analyzed using a thematic analysis approach. The second part involved document inspection. Each participant was first asked to submit one of their Word documents or PowerPoint presentations (source file) and the corresponding PDF file. Then they were asked to make the documents (both sources and PDF files) accessible using the information provided by the interactive guide and submit the accessible source files and the accessible PDF files. The Word and PowerPoint files were analyzed using the Accessibility Checker built into Microsoft 365 and manual checks. PDF files were analyzed using the Accessibility Checker in Adobe Acrobat Pro, and the command-line exiftool for checking the meta-data of the PDF files. At the end of each part of the evaluation, the participants were asked to complete the System Usability Scale (SUS) questionnaire (Brooke, 1995) which includes ten statements adapted to the interactive guide. The study was reported to the Data Protection Services in the National Agency for Shared Services in Education and Research (Sikt) which evaluated and approved the processing of personal data in this study.

### 4.2 Results

#### 4.2.1 Participants

The participants were recruited using a convenience sampling method. Emails were sent to colleagues in the university and in our professional networks. Six participants (P1-P6) took
part in the usability testing and eight (P7-P14) took part in the document inspection. All participants completed the SUS questionnaire. The working areas of the participants are shown in Table 1.

Table 1. The working areas of the participants

<table>
<thead>
<tr>
<th>Position type</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>P1, P2, P4, P5, P12</td>
</tr>
<tr>
<td>Teaching and Research</td>
<td>P3, P6, P7, P8, P10, P11, P13, P14</td>
</tr>
<tr>
<td>Library</td>
<td>P9</td>
</tr>
</tbody>
</table>

4.2.2 Document Inspection

Seven participants submitted the before and after versions of Word documents and PDF files and one participant submitted the before and after version of a PowerPoint presentation. The accessibility inspection of the submitted Word documents, PowerPoint presentations and the corresponding PDF file focused on the components including title, language, heading, text, image, table, link, reading order and slide title for PowerPoint presentations and language and bookmark in PDF files.

Table 2 presents the results from the document inspection. Because the prototype of the interactive guide has not yet provided all the guidance concerning the components, in Table 2 we only include those that the guide has covered. The document inspection showed that participants were able to improve their documents in text alternatives of images, color contrast, and the correct use of headings. Accessibility issues have also been found in document titles, language (both in source and PDF files) and PDF bookmarks. Most of the accessibility issues found with tables were not corrected after the participants used the interactive guide. One possible explanation could be that some of those issues are not covered by the built-in accessibility checker in Microsoft 365, which is the case for tables missing heading rows. The automatic checker did report tables with split or merged cells, but those errors were not fixed by the participants in this study. It could be that this topic needed to be addressed better in the interactive guide or that the participants needed more motivation or time to correct them.

Table 2. Accessibility issues found in document inspection (before/after)

<table>
<thead>
<tr>
<th>Components</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heading</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>Text*</td>
<td>136</td>
<td>36</td>
</tr>
<tr>
<td>Image*</td>
<td>96</td>
<td>65</td>
</tr>
<tr>
<td>Table*</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>Link*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reading order (PowerPoint)</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Slide title (PowerPoint)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: components with * mean that the interactive guide does not cover all the accessibility issues, and the number in the table for these components shows only the covered issues.

4.2.3 SUS Score

The lowest SUS score is 42.5 given by P9 and the highest is 82.5 given by P12. The mean score is 66.1 and standard deviation is 15.2. According to Bangor et al. (2009), if a SUS mean score is over 60 the system evaluated can be considered “Good”, and only when the mean score is over 75 can the system be considered “Excellent”. In our study, the mean score is 66.1, so it has only reached the “Good” level.
4.2.4 Results from Qualitative Analysis

The qualitative data is collected through observations and semi-structured interviews associated with user testing and the open-ended questions after the SUS questionnaire. The participants generally thought the interactive guide was clear and easy to use. We observed that the participants had different preferences concerning using the table of content or the filters. The table of contents gave an overview of the guides, including the important points for users to remember when creating accessible content. Some participants preferred the filtering function while others felt that they did not need the filter options because the table of contents is not very long. The participants have also reported that they have learned new techniques from the online guide, the websites linked from the online guide and through discussions during the user testing sessions. For example, one participant was asked to create an accessible table in the user testing. The participant carried out the task as s/he used to do. After being encouraged to look at the online guide and the linked video from the guide, the participant was able to create an accessible table. Soon after the testing, the participant needed to prepare a large report with many tables and was happy to report back that s/he was able to put the new skills to use.

The participants have also provided valuable feedback to improve the online guide concerning both the usability and the content of the guide. Some suggested more content, including more details related to image text alternatives, working with tables in an HTML source editor in an LMS, and captioning videos. Several participants suggested the guide should be more practically oriented and use images, short videos or more advanced HTML code examples to demonstrate the techniques.

For people who would not have the expertise, […] it's difficult for them to understand what it means without pictures or images that should demonstrate how things should be done. […] if it's a user guide you have pictures to guide the users on how to do things. (P9).

Concerning the organization of the content in the online guide, one of the participants suggested that the topics should be organized in different ways (e.g. based on software or applications) so that they satisfy the needs of users with different preferences.

5. Discussion

Studies have shown that students with disabilities face a series of barriers that hinder their success in higher education and faculty members have been identified as one of the main barriers for students with disabilities in higher education (Zhang et al., 2018). Research has also shown that faculty members lack motivation and relevant knowledge and skills to respond to the needs of students with disabilities. Staff training and administrative support have been recommended as measures to address the barriers related to faculty members.

Several studies (Chen et al., 2018; Linder et al., 2015; Sanderson et al., 2022) have argued that in order to provide training and support to faculty members, higher education institutions should establish policies and strategies, and allocate resources for implementation. Some participants in our study have also commented on the responsibilities of institutions and pointed out that faculty members should not be the only responsible party. In our study, the participants are motivated to learn and make their digital content accessible. They are aware that they lack the necessary knowledge and skills and expressed that they wish to have someone they can ask when facing challenges with the accessibility of their digital content or get help about what tools are good and how to use them. They thought that this online guide was timely and important.

The online guide does not cover all possible accessibility issues. The accessibility improvements the participants made in their files when using the online guide were not very impressive, but there were some improvements. Some participants had yet to improve on all the issues the guide covered after using it. This could partly be explained by how motivated the participants were and how much time they were willing to spend testing the guide and improving their files in a busy life. Nevertheless, the data suggest that the online guide has potential but needs further improvement.
6. Conclusion and Future Work

The research presented in this paper aims to address the challenges faced by content creators in higher education in creating accessible digital content by providing an easy-to-use online guide. The online guide includes carefully selected content which can help content creators with the task at hand so that they do not have to search through a large amount of online information. Despite some usability challenges, the participants were able to use the guide to learn new techniques and improve the accessibility of their digital contents. The results suggest that there is potential in an online guide such as this to help content creators in creating accessible digital content.

References


