

Educational Technology & Society

Published by International Forum of Educational Technology & Society
Hosted by National Taiwan Normal University, Taiwan

Jan. 2023

Educational Technology & Society

Educational Technology & Society has Impact Factor 2.633 and
5-Year impact factor 4.358 according to Thomson Scientific 2021 Journal Citations Report.

<http://www.j-ets.net/>

vol. **26**
no. **1**

Volume **26** Issue **1**

January 2023

ISSN: 1436-4522 (online)
ISSN: 1176-3647 (print)
DOI: 10.30191/ETS
<http://www.j-ets.net/>

Educational Technology & Society

An International Journal

Aims and Scope

Educational Technology & Society (ET&S) is an open-access academic journal published quarterly (January, April, July, and October) since October 1998. *ET&S* has achieved its purposes of providing an international forum for open access scientific dialogue for developers, educators and researchers to foster the development of research in educational technology. Thanks to all the Authors, Reviewers and Readers, the journal has enjoyed tremendous success.

ET&S has established a solid and stable editorial office where the Editors-in-Chief aims to promote innovative educational technology research based on empirical inquiries to echo the pedagogical essentials of learning in the real world—lifelong learning, competency-orientation, and multimodal literacy in the 21st century.

ET&S publishes research that well bridges the pedagogy and practice in advanced technology for evidence-based and meaningful educational application. The focus of *ET&S* is not only technology per se, but rather issues related to the process continuum of learning, teaching, and assessment and how they are affected or enhanced using technologies. The empirical research about how technology can be used to overcome the existing problems in the frontline of local education with findings that can be applied to the global spectrum is also welcome. However, papers with only descriptions of the results obtained from self-report surveys without systematic or empirical data or any analysis on learning outcomes or processes are not favorable to be included in *ET&S*.

ET&S publishes research that fulfills any of the following description:

- **Evidence-based Research:** Continuous research that is conducted within a sufficient amount of time such that the effectiveness of the research intervention in education can be evaluated and demonstrated convincingly via empirical methods. Preferably, research experience, outcome, and results for a semester long of adopting educational technologies; or an evaluation/experiment that is conducted multiple times and spans over several semesters.
- **Emerging Technology:** New and emerging technology used in education
- **Special Domain Research:** Research with specific participants using technology in education, for example, those with privacy issues in special education or those of younger ages like infants and toddlers, who may not form a large pool of participants

Founding Editor

Kinshuk, University of North Texas, USA.

Journal Steering Board

Nian-Shing Chen, National Taiwan Normal University, Taiwan; **Kinshuk**, University of North Texas, USA; **Demetrios G. Sampson**, University of Piraeus, Greece.

Editors-in-Chief

Maiga Chang, Athabasca University, Canada; **Dirk Ifenthaler**, University of Mannheim, Germany; **Yu-Ju Lan**, National Taiwan Normal University, Taiwan.

Associate Editors

Yacine Atif, University of Skövde, Sweden; **Howard Hao-Jan Chen**, Department of English, National Taiwan Normal University, Taiwan; **Jun (Scott) Chen Hsieh**, Asia University, Taiwan; **Yannis Dimitriadis**, Universidad de Valladolid, Spain; **Kyparisia Papanikolaou**, School of Pedagogical & Technological Education, Greece; **Yun Wen**, National Institute of Education, Singapore.

Editorial Board Members

Ahmed Hosny Saleh Metwally, Northeast Normal University, China; **Bernardo Pereira Nunes**, The Australian National University, Australia; **Ching-sing Chai**, The Chinese University of Hong Kong, Hong Kong; **David Gibson**, Curtin University, Australia; **Grace Yue Qi**, Massey University, New Zealand; **Ig Ibert Bittencourt Santana Pinto**, Universidade Federal de Alagoas, Brazil; **Jerry Chih-Yuan Sun**, National Chiao Tung University, Taiwan; **Jie Chi Yang**, National Central University, Taiwan; **Joice Lee Otsuka**, Federal University of São Carlos, Brazil; **Kaushal Kumar Bhagat**, Indian Institute of Technology, India; **Minhong Wang**, The University of Hong Kong, Hong Kong; **Morris Siu-Yung Jong**, The Chinese University of Hong Kong, Hong Kong; **Regina Kaplan-Rakowski**, University of North Texas, USA; **Rita Kuo**, New Mexico Tech, USA; **Robert Li-Wei Hsu**, National Kaohsiung University of Hospitality and Tourism, Taiwan; **Rustam Shadiey**, Nanjing Normal University, China; **Stephen J.H. Yang**, National Central University, Taiwan; **Tony Liao**, NOAA Earth System Research Laboratories, USA; **Wen-Ta Tseng**, National Taiwan University of Science and Technology, Taiwan; **Yanjie Song**, Education University of Hong Kong, Hong Kong; **Ahmed Tlili**, Smart Learning Institute of Beijing Normal University, China; **Chia-Wen Tsai**, Department of Information Management, Ming Chuan University, Taiwan; **Hsueh Chu Rebecca Chen**, Education University of Hong Kong, Hong Kong; **Nur Hamid**, Universitas Islam Negeri Walisongo Semarang, Indonesia; **Katrin Saks**, University of Tartu, Estonia; **Sheng-Shiang Tseng**, Tamkang University, Taiwan; **Siddharth Srivastava**, Indian Institute of Technology Kanpur, India; **Li Wang**, Open University of China, China; **Robin Jocius**, University of Texas at Arlington, USA; **Dongsik Kim**, Hanyang University, South Korea; **Chiu-Lin Lai**, National Taipei University of Education, Taiwan; **Daner Sun**, Education University of Hong Kong; **Ying-Tien Wu**, National Central University, Taiwan; **Hui-Chin Yeh**, National Yunlin University of Science and Technology, Taiwan.

Managing Editor

Sie Wai (Sylvia) Chew, National Taiwan Normal University, Taiwan.

Editorial Assistant

I-Chen Huang, National Taiwan Normal University, Taiwan.

Technical Manager

Wei-Lun Chang, National Taiwan Normal University, Taiwan.

Executive Peer-Reviewers

see <http://www.j-ets.net>

Publisher

International Forum of Educational Technology & Society

Host

National Taiwan Normal University, Taiwan

Editorial Office

c/o Chair Professor Nian-Shing Chen, Institute for Research Excellence in Learning Sciences, Program of Learning Sciences, National Taiwan Normal University, No.162, Sec. 1, Heping E. Rd., Da-an Dist., Taipei City 10610, Taiwan.

Supporting Organizations

University of North Texas, USA

University of Piraeus, Greece

National Yunlin University of Science and Technology, Taiwan

National Sun Yat-sen University, Taiwan

Website Maintenance

Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan

Abstracting and Indexing

Educational Technology & Society is abstracted/indexed in Social Science Citation Index, Scopus, ACM Guide to Computing Literature, airiti, Australian DEST Register of Refereed Journals, Computing Reviews, Current Contents/Social & Behavioral Sciences, DBLP, DOAJ, Educational Administration Abstracts, Educational Research Abstracts, Educational Technology Abstracts, Elsevier Bibliographic Databases, ERIC Clearinghouse on Information & Technology, Inspec, ISI Alerting Services, JSTOR, PsycINFO, Social Scisearch, Technical Education & Training Abstracts, and VOCED.

Guidelines for authors

Submissions are invited in the following categories:

- Peer reviewed publications: Full length articles (up to 8,000 words excluding References and Appendices)
- Special Issue publications

All peer review publications will be refereed in double-blind review process by at least two international reviewers with expertise in the relevant subject area.

For detailed information on how to format your submissions, please see: https://www.j-ets.net/author_guide

For Special Issue Proposal submission, please see: https://www.j-ets.net/journal_info/special-issue-proposals

Submission procedure

All submissions must be uploaded through our online management system (<http://www.j-ets.net>). Do note that all manuscripts must comply with requirements stated in the Authors Guidelines.

Authors, submitting articles for a particular special issue, should send their submissions according to the channel specified in the Call for Paper of the special issue.

All submissions should be in electronic form. Authors will receive an email acknowledgement of their submission.

The preferred formats for submission are Word document, and not in any other word-processing or desktop-publishing formats. Please place figures and tables in their respective format in the anonymous manuscript along with all appendices (if any).

Please provide following details with each submission in a separate file (i.e., Title Page):

- Author(s) full name(s) including title(s),
- Name of corresponding author,
- Job title(s),
- Organisation(s),
- Full contact details of ALL authors including email address, postal address, telephone and fax numbers.

In case of difficulties, please contact journal.ets@gmail.com.

Table of Contents

Full Length Articles

- Students' Social-Cognitive Engagement in Online Discussions: An Integrated Analysis Perspective 1–15
Zhi Liu, Ning Zhang, Xian Peng, Sannyuya Liu and Zongkai Yang
- Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness 16–30
Siu-Cheung Kong, William Man-Yin Cheung and Guo Zhang
- A Systematic Review of Technology-Enhanced Self-Regulated Language Learning 31–44
Yin Yang, Yun Wen and Yanjie Song
- Exploring the Research Trajectory of Digital Game-based Learning: A Citation Network Analysis 45–61
Wiwit Ratnasari, Tzu-Chuan Chou and Chen-Hao Huang
- Influences of Growth Mindset, Fixed Mindset, Grit, and Self-determination on Self-efficacy in Game-based Creativity Learning 62–78
Yu-chu Yeh, Yu-Shan Ting and Jui-Ling Chiang
- Semiotic Alternations with the Yupana Inca Tawa Pukllay in the Gamified Learning of Numbers at a Rural Peruvian School 79–94
Rosario Guzman-Jimenez, Dhavit-Prem, Alvaro Saldívar and Alejandro Escotto-Córdova

Editorial Note

- Human-centered AI in Education: Augment Human Intelligence with Machine Intelligence 95–98
Stephen J.H. Yang, Hiroaki Ogata and Tatsunori Matsui

Special Issue Articles

- Unpacking the “Black Box” of AI in Education 99–111
Nabeel Gillani, Rebecca Eynon, Catherine Chiabaut and Kelsey Finkel
- Trends, Research Issues and Applications of Artificial Intelligence in Language Education 112–131
Xinyi Huang, Di Zou, Gary Cheng, Xiuling Chen and Haoran Xie
- A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior 132–146
Fuzheng Zhao, Gi-Zen Liu, Juan Zhou and Chengjiu YinBiyun
- A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning 147–159
Alwyn Vwen Yen Lee, Andrés Carlos Luco and Seng Chee Tan
- Feasibility and Accessibility of Human-centered AI-based Simulation System for Improving the Occupational Safety of Clinical Workplace 160–170
Pin-Hsuan Wang, Anna YuQing Huang, Yen-Hsun Huang, Ying-Ying Yang, Jiing-Feng Lirng, Tzu-Hao Li, Ming-Chih Hou, Chen-Huan Chen, Albert ChihChieh Yang, Chi-Hung Lin and Wayne Huey-Herng Sheu
- Artificial Intelligent Robots for Precision Education: A Topic Modeling-Based Bibliometric Analysis 171–186
Xiuling Chen, Gary Cheng, Di Zou, Baichang Zhong and Haoran Xie
- A Risk Framework for Human-centered Artificial Intelligence in Education: Based on Literature Review and Delphi–AHP Method 187–202
Shijin Li and Xiaoqing Gu

AI, Please Help Me Choose a Course: Building a Personalized Hybrid Course Recommendation System to Assist Students in Choosing Courses Adaptively <i>Hui-Tzu Chang, Chia-Yu Lin, Wei-Bin Jheng, Shih-Hsu Chen, Hsien-Hua Wu, Fang-Ching Tseng and Li-Chun Wang</i>	203–217
Effects of Incorporating an Expert Decision-making Mechanism into Chatbots on Students' Achievement, Enjoyment, and Anxiety <i>Ting-Chia Hsu, Hsiu-Ling Huang, Gwo-Jen Hwang and Mu-Sheng Chen</i>	218–231
Application of Artificial Intelligence Techniques in Analysis and Assessment of Digital Competence in University Courses <i>Tzu-Chi Yang</i>	232–243

Students' Social-Cognitive Engagement in Online Discussions: An Integrated Analysis Perspective

Zhi Liu¹, Ning Zhang², Xian Peng^{1*}, Sannyuya Liu^{1,2} and Zongkai Yang^{1,2}

¹National Engineering Research Center for Educational Big Data, Faculty of Artificial Intelligence in Education, Central China Normal University, China // ²National Engineering Research Center for E-Learning, Faculty of Artificial Intelligence in Education, Central China Normal University, China // zhiliu@mail.ccnu.edu.cn // zhangning97@mails.ccnu.edu.cn // pengxian@ccnu.edu.cn // lsy.nercel@gmail.com // 13659885363@163.com
*Corresponding author

(Submitted November 23, 2021; Revised February 11, 2022; Accepted March 18, 2022)

ABSTRACT: Grounded on constructivism, mining a complex mix of social and cognitive interrelations is key to understanding collaborative discussion in online learning. A single examination of one of these factors tends to overlook the impact of the other factor on learning. In this paper, we innovatively constructed a social-cognitive engagement setting to jointly characterize social and cognitive aspects. In the online discussion forum, this study jointly characterized students' social and cognitive aspects to investigate interactive patterns of different social-cognitive engagements and social-cognitive engagement evolution across four periods (i.e., creation, growth, maturity, and death). Multi-methods including social network analysis, content analysis, epistemic network analysis, and statistical analysis was applied in this study. The results showed that the interactive patterns of social-cognitive engagement were affected by both social network position and cognitive level. In particular, students' social network position was a vital indicator for the contributions to cognitive level of students, and cognitive level affected the related interactions to some extent. In addition, this study found a nonlinear evolutionary development of students' social-cognitive engagement. Furthermore, maturity is a critical period on which teachers should focus, as the co-occurrence of social-cognitive engagement reaches a maximum level in this period. Based on the results, this multi-perspective analysis including social and cognitive aspects can provide insightful methodological implications and practical suggestions for teachers in conducting in-depth interactive discussions.

Keywords: Social-cognitive engagement, Integrated analysis, Social network analysis (SNA), Epistemic network analysis (ENA), Knowledge building

1. Introduction

With advances in Internet technology and the large-scale application of computer-mediated communication tools, online courses have experienced tremendous growth. As the primary space for students to have discussions by posting messages, the online discussion forum provides participants with the support to interact with their peers or instructors as well as various materials for learning. Existing studies have confirmed that online discussions facilitate knowledge construction and learning engagement (Cukurova et al., 2018).

Social and cognitive aspects are two important factors that can affect academic performance in terms of parsing the interactive process of student discourse discussions (Liu & Matthews, 2005). The social aspect mainly refers to social interaction, participation, and perspective taking (Hesse et al., 2015). The cognitive aspect typically concerns knowledge construction, cognitive inquiry, and problem solving (Ouyang & Chang, 2019; Swiecki & Shaffer, 2020). According to constructivism theory (Liu & Matthews, 2005), capturing the complex interactions between the interrelated social and cognitive aspects is essential for demonstrating collaborative discussion in online learning.

Recently, several researchers have begun to attempt a joint analysis of students' social and cognitive aspects in online forums. Some studies are devoted to using multiple methods to investigate social and cognitive aspects (Peng & Xu, 2020) and their interrelationship (Tirado et al., 2015); some studies focus on the design and updating of the research framework (Ke & Xie, 2009; Wang et al., 2014); and other studies pay attention to the proposal and application of new methods (Gašević et al., 2019; Swiecki & Shaffer, 2020).

Although the above studies have provided insights into the integrated analysis of social and cognitive aspects, they have considered these aspects separately, rather than attempting to view them as a whole, which ignores the joint contribution of both to discussion-based learning. To address this issue, our research provides researchers with a novel perspective of refining students' social and cognitive discourse characteristics. Specifically, this

study extracted the social and cognitive aspects of the students' characteristics separately and combined them into a new characteristic—social-cognitive engagement. Social-cognitive engagement is not a simple sum of social and cognitive characteristics, but a cross-combination that reflects both social and cognitive aspects. This cross-combination considers the intertwined impact of one aspect on the other aspect in the online discussion forum. Moreover, unlike a traditional network in which nodes represent a single attribute, the construction of social-cognitive engagement allows for the visualization of the social and cognitive aspects of nodes for a fine-grained characterization of node states and node relationships in the network.

Multi-methods were addressed to investigate the relationship between social and cognitive aspects in online discussions and interactive patterns of different social-cognitive engagements, as well as social-cognitive engagement evolution. Results can be used to uncover the evolutionary patterns of students in the learning process and help teachers better understand students' knowledge construction process in detail so that they can design reasonable teaching plans.

2. Theoretical framework

2.1. Theoretical foundations

Sfard (1998) explained learning as two metaphors: acquisition and participation. The acquisition metaphor demonstrates that learning can be understood through the acquisition of knowledge by individuals in their minds, while the participation metaphor suggests that learning is facilitated through social interactions between individuals in a community of practice (Teo et al., 2017). Combining the acquisition metaphor and the participation metaphor, Paavola et al. (2004) proposed a third metaphor: knowledge creation. Knowledge creation assumes that individuals engage in collaborative discussions within a community, acquiring personal knowledge and creating new knowledge that can be used in the whole community. As the main model supporting the conceptualization of knowledge creation communities, knowledge building theory makes the argument for “learning as knowledge creation” explicit and well-documented.

Knowledge building is based on the theoretical guidance of constructivism (Bereiter, 2002b; Yücel & Usuel, 2016) and can be defined as a learning process in which students generate different ideas and develop, integrate, refine, or elaborate these ideas through progressive discussion activities (Lin et al., 2014). Knowledge building involves not only students sharing their ideas but also further negotiation and discussion based on existing views and thoughts. Therefore, knowledge building stresses the social interaction process of the formation of a knowledge community.

Figure 1. Theoretical foundations of social-cognitive engagement



Hong et al. (2010) proposed that the principles guiding knowledge building could be summarized from three dimensions: ideas, agents, and community. Ideas are the building blocks of knowledge, and idea improvement is an extremely important part of knowledge building. Sometimes, ideas proposed by students may not be correct or lack reasonable explanations, but students' ideas can be changed through social interaction. This process is a major contribution to constructive learning; therefore, ideas are referred to as epistemic anchors (Bereiter, 2002a). Agents are knowledge workers who are treated as the subjects of knowledge. They obtain knowledge through sustained idea improvement and collaborative learning patterns. Agents need to assume epistemic agency and engage in the constructive use of authoritative activities. Community is a social venue for knowledge or idea interactions.

According to Descartes (Lin et al., 2014), the two most fundamental epistemological aspects concern with the object of learning, i.e., what people want to know, and the subject of learning, i.e., the learner. Between the two aspects there is also a social aspect that defines the social space in which learning takes place. The above three

epistemological aspects (objective, subjective and social) constitute the conceptual framework for knowledge building. Accordingly, in this study, the three dimensions of knowledge building were mapped to the three objects of the online engagement, i.e., cognitive aspects, students, and social aspects. As shown in Figure 1. Students engage in interactions in community and generate different social characteristics indicated by the frequency, direction, and object of the interaction. In addition, their knowledge or ideas reflect different cognitive levels. Based on the constructivism, this study proposed the concept of social-cognitive engagement in an attempt to portray the learning process in online discussion. The social-cognitive engagement construction process was described in section 3.4 (Data analysis).

2.2. Related literature

Social network analysis (SNA) and content analysis (CA) are commonly used to analyze the social and cognitive aspects in discussion-based learning. Recently, epistemic network analysis (ENA) has gained a lot of attention as a network analysis method to model interactions among cognitive elements (Shaffer et al., 2016). Existing literature demonstrated that ENA can be widely used in the field of learning analytics, such as thinking development (Tan et al., 2022), learning evaluation (Fougt et al., 2018) and knowledge construction (Shaffer et al., 2016). For example, Bressler et al. (2019) used ENA to examine the evolution of collaborative scientific practice and discourse of student team. Tan et al., (2022) explored the development trajectory of shared epistemic agency in collaborative learning through ENA. Overall, researchers can identify changes in students' cognitive development through comparative analysis of their cognitive networks at different periods with the help of ENA, which is conducive to guiding and nurturing students' cognitive development in actual teaching activities.

Several studies have attempted to combine SNA and CA to investigate the relationship between social and cognitive aspects (Tirado et al., 2015; Zhang et al., 2017). The coding frameworks have also been redesigned to integrate these two aspects. Social presence and cognitive presence in the community of inquiry framework (CoI) have been used to describe social and cognitive issues in learning community (Garrison et al., 2010; Popescu & Badea, 2020). An online learning interaction model was developed by Ke and Xie (2009) to evaluate the objective evidence of adults' social and cognitive engagement. The framework addresses social interactions, knowledge construction processes, and self-directed processes. Wang et al. (2014) constructed a framework for interaction and cognitive engagement in connectivist learning contexts. Operation, wayfinding, sensemaking, and innovation comprised the four levels of the framework. In addition to the updating and designing of the framework, some researchers have proposed new approaches to integrating social and cognitive aspects for analysis. The social-epistemic network signature (SENS) is a network analytics approach combining the social and cognitive perspectives of collaborative learning (Gašević et al., 2019) and the use of SENS proves that cognitive and social aspects can be modeled as networked. Also, combining SNA and ENA, Swiecki and Shaffer (2020) proposed an integrated social-epistemic network signature (iSENS).

All of the above works can be described as integrated studies on social and cognitive aspects. Nevertheless, these studies still narrate the findings separately in terms of social and cognitive aspects. For example, studies combining SNA and CA often perform correlation analysis between different network measures identified by SNA and various cognitive behaviors encoded by CA (Ouyang & Chang, 2019; Zhang et al., 2017). Coding frameworks tend to define social and cognitive aspects as two different dimensions (Ke & Xie, 2009) or artificially equate high levels of social aspect with high levels of cognitive aspect (Wang et al., 2014). Even some new methods, such as SENS, regard social and cognitive patterns as independent predictors (Swiecki & Shaffer, 2020). In summary, previous studies still separate social and cognitive aspects. Unlike previous studies, this study used SNA and CA to construct students' social-cognitive engagement, and thus explored the interactive pattern of different social-cognitive engagements. In addition, this study was interested in examining the evolution of social-cognitive engagement since learning is a dynamic evolutionary process of acquiring knowledge. Nowadays, some studies have started to focus on the dynamic evolution of online forums, such as the evolution of topic content over the duration (Peng et al., 2020; Peng & Xu, 2020) and the comparison of social networks for multi-round activity (Zhang et al., 2017). Grasping the evolutionary trends of students' dynamistic interactions can help us uncover the evolutionary patterns of students in the learning process.

2.3. Research questions

Although few studies have attempted an integrated analysis of social and cognitive aspects, the social and cognitive aspects remain fragmented, which prevents us from gaining a deeper understanding of the social and cognitive connections and their joint impact on online forums. Moreover, the interactive process of student

discourse discussions cannot be understood without examining the social and cognitive aspects embodied in interactions. To tackle these issues, this study tracked students' social-cognitive engagement on online forums from the perspective of joint modeling. Specifically, this study aimed to address the following research questions:

- RQ1: What is the relationship between social and cognitive aspects in online discussions?
- RQ2: What is the pattern of students' social-cognitive engagement in online discussions?
- RQ3: How does students' social-cognitive engagement evolve at the different phases of online discussions?

3. Methodology

3.1. Research context

The forum data were collected from a learning platform developed by National Engineering Research Center for E-Learning of Central China Normal University (CCNU). The platform in this study embeds multiple learning resources (e.g., videos, courseware, and quizzes) and learning contexts (e.g., group discussions and independent learning).

Before the course, all participants had been trained to use the platform. According to the teaching schedule, a teaching assistant posted the discussion topics on the platform every week, and students participated in the discussions. Neither the teaching assistant nor the teacher interfered with student discussions throughout the teaching activities. Fourteen discussion topics centered on the course "Introduction to Data Science" were designed to deepen these students' knowledge and understanding of data science. The discussion topics on the course content were initially delivered by the teaching assistant before the start of each class. The discussion topics covered data visualization, correlation analysis, and so on. Students could find these topics on the platform and participate in discussions, as shown in Figure 2. It is worth noting that China went through its largest online learning period in 2020 due to the COVID-19 pandemic, and this course was conducted in the first full semester after students had returned to school.

Figure 2. Screenshot of the discussion forum in a MOOC platform

The screenshot shows a MOOC platform interface. At the top, there are navigation links: 'my note', '我的笔记', '我的课程', and 'my course'. Below this, a sidebar on the left contains various navigation options: '数据科学导论', '课程主页 → course page', '课程内容 → course content', '任务 → task', '成员 → members', '分组 → group', '成绩 → achievement', and '课程画像 → course profile'. The main content area displays a list of discussion topics. A red arrow points to the '任务' (task) section in the sidebar, with the label 'different discussion topics' next to it. The list of topics includes titles, task types (all '讨论' or 'discussion'), status (all '已截止' or 'closed'), start times, and end times.

标题	任务类型	状态	开始时间	截止时间
【发布了《我们正处在大数据时代中，有没有哪些地方让你感受到“大数据”的存在？请举例说明。另外，数据与科学是如何结合成为数据科学的？教育领域中是否也有大数据之说？》在【第二讲 数据科学与大数据】	讨论	已截止	2020-10-03 11:05:09	2020-10-12 20:00:00
【发布了《你听到“大数据”这个词是什么时候？当看到“数据科学导论”这门课时，你的感受是什么？同时从自身角度出发，谈谈你对数据科学的理解。》在【第一讲 数据科学概述】	讨论	已截止	2020-09-28 09:00:17	2020-10-05 20:00:17
【发布了《请结合自身理解谈谈什么是数据可视化，以及为什么将数据进行可视化？你知道有哪些可以将数据实现可视化的工具，它们各自有什么优缺点？另外，作为一名数据可视化工程师，你认为应该掌握哪些技能。》在【第十二讲 数据可视化】	讨论	已截止	2020-12-11 15:13:30	2020-12-17 15:13:30
【发布了《通过查找资料及阅读文献，请详细介绍一个智能导学系统(国内外皆可)，如果可以，你希望在这	讨论	已截止	2020-12-18 17:58:49	2020-12-25 17:58:49

3.2. Participants

The participants of this study were 35 undergraduate students who attended the course "Introduction to Data Science" at a university in Wuhan, China. All of them majored in data science and big data technology. The final course grades consisted of weighted scores from ordinary grades (70%) and final examination scores (30%).

Ordinary grades included online learning hours, attendance, collaborative activity participation, and online participation in the discussion forum. For personal reasons, one student did not participate in the final examination and was therefore excluded from the analysis related to academic performance. Final grades were normalized on a 0-100 scale ($N = 34$, $M = 83.45$, $SD = 5.40$). In total, 1,068 messages were posted by 35 students.

3.3. Measures

Revised Bloom's Taxonomy (RBT) was adopted as the coding scheme to operationalize cognitive behaviors to better capture students' cognitive aspects in this study. *Remember*, *understand*, and *apply* are defined as lower-order cognitive behaviors while *analyze*, *evaluate*, and *create* are defined as higher-order cognitive behaviors. In addition, *off-topic* was added to indicate the student discussions that were irrelevant to the course content, as denoted in Table 1.

Table 1. Coding scheme for cognitive behavior

Code	Categories	Example
B1	Remember	Big data is data collection that cannot be crawled, managed...
B2	Understand	The amount of data generated in the era of Big Data is incomparable to any previous period in human history ... and it will be a challenge to store such a large amount of information.
B3	Apply	I would like to know what you think are the practical problems in education and...
B4	Analyze	Artificial intelligence cannot be separated from the support of big data, because ... Deep learning is a new development direction in machine learning...
B5	Evaluate	The added system you mentioned at the end is pretty novel...and having the smart guide system simulate the idea of being a student.
B6	Create	Although Auto Tutor already implements...I would like to add a "deep personalization system" to Auto Tutor...By studying the interpersonal interactions...
B7	Off-topic	After studying, we all have a new understanding of this course, so let's do it together!

Two experienced researchers who were familiar with the RBT were invited to jointly code all 1,068 discussion messages manually. To ensure the reliability of the coding results, the kappa value was calculated, and that of the two coders was 0.76, indicating that the coded results were reliable. For other inconsistent codes, the two coders had several discussions until a consistent result was obtained.

In this study, Hyperlink-Induced Topic Search (HITS) and PageRank were provided to measure the importance of nodes in a network. PageRank is a measure for scoring the importance of nodes based on the linking relationships between them. Hub and authority are the measures of HITS. A good hub is usually linked to many other nodes and a good authority is usually linked by various hubs.

3.4. Data analysis

To address the three research questions raised, a series of methods, such as CA, SNA, ENA, and statistical methods, was applied in this study.

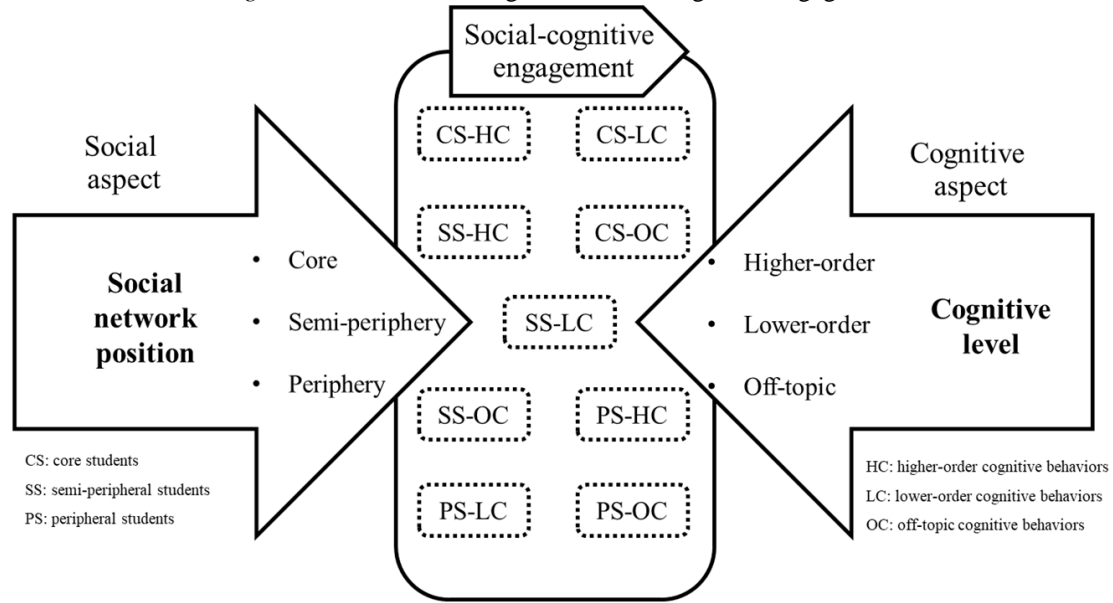
To answer the first research question, cognitive behaviors were coded, and coreness was calculated to identify students in different social network positions. Kruskal–Wallis nonparametric tests were employed to determine if there were statistically significant differences in cognitive behavior among different categories of students. When the Kruskal–Wallis tests prove a significant difference, post hoc tests (Mann–Whitney U Tests) should be performed to determine which two categories of students differ. In addition, ENA was employed to characterize the epistemic network of different categories of students.

To answer the second research question, social-cognitive engagement was constructed based on the classification of network position and cognitive level. An interactive network diagram of social-cognitive engagement was depicted to demonstrate its interactive pattern. Moreover, social network measures were calculated to examine the importance of each social-cognitive engagement in interactive networks. In addition, the Kruskal–Wallis nonparametric test and Mann–Whitney U test were used to compare the differences in academic performance in students' social-cognitive engagement.

To answer the third research question, this study divided the learning process into six phases. After referring to other similar literatures as well as conducting the actual analysis of this study, we found that when the forum was divided into six phases, on average, each student interacted with others more than 2 times per phase. This means that students will interact with more than one person or interact with the same person more than once, which can contribute to mining the evolutionary characteristics of the network as much as possible. As a measure of stability, the Jaccard coefficients (Huang et al., 2021; Zhang et al., 2016) for two sequential phases varied from 0.620 to 0.915, indicating that the network dynamics of these six phases were smooth enough and appropriate for this study. Moreover, this study employed ENA to detect the evolution of social-cognitive engagement during the discussions.

In the process of data analysis, there is one key point worth addressing: How can social-cognitive engagement be constructed? The specific analytical process was elaborated in detail as follows.

Figure 3. Construction diagram of social-cognitive engagement



In the social-cognitive engagement proposed in this study, both social and cognitive aspects should be demonstrated. On the one hand, among the two core/periphery structures proposed by Borgatti and Everett (1999), the continuous core/periphery structure uses “coreness,” a quantitative indicator of network position, to determine the relative position of each node in the network and to divide the core and periphery sets of the social network. Students can be classified as core students, semi-peripheral students, or peripheral students according to their coreness. On the other hand, students’ cognitive behaviors can be divided into higher-order cognitive behaviors, lower-order cognitive behaviors, and off-topic cognitive behavior. Based on social network position and cognitive level, this study constructed nine social-cognitive engagements: CS-HC, CS-LC, CS-OC, SS-HC, SS-LC, SS-OC, PS-HC, PS-LC, and PS-OC. Figure 3 illustrates the construction process. Consider the following sentence as an example: “After studying, we all have a new understanding of this course, so let’s do it together!” The sentence was proposed by a student whose ID was S2. This student was defined as a core student. Additionally, the sentence was coded as off-topic. In combination, the social-cognitive engagement of this sentence was labeled CS-OC, which denotes post reflecting *off-topic* cognitive behavior proposed by a core student.

4. Results

4.1. Relationship between social and cognitive aspects in online discussions

To answer RQ1, we calculated coreness to classify students in different social network positions and coded cognitive behaviors within an online discussion, respectively. Non-parametric tests were used to identify the differences in the cognitive behavior of different categories of students, and ENA was employed to characterize the epistemic network of different categories of students.

After calculating, the correlation between the data and the idealized core/periphery structure was 0.917, indicating a good fit of the core/periphery model. The Gini coefficient was 0.641, suggesting that the coreness varied greatly among the nodes. According to the principle of classifying students with coreness greater than or equal to 0.2 as core students and those with coreness lower than or equal to 0.05 as peripheral students, the remaining students were defined as semi-peripheral students. Figure 4 shows the social network diagram of online discussion. The individual students within interactive networks were represented as nodes, and interactive relationships were visualized with lines between the nodes. In Figure 4, there are five core students (green nodes) surrounded by semi-peripheral students (orange nodes, $N = 18$) and peripheral students (purple nodes, $N = 12$).

Figure 4. The social network diagram of online discussion

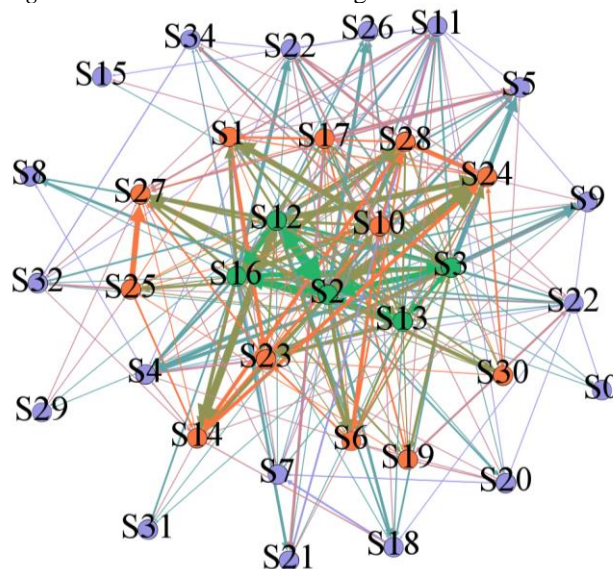


Table 2. Differences in the cognitive behaviors of different students

Type	Core students (3, $n = 5$)	Semi-peripheral students (2, $n = 12$)	peripheral students (1, $n = 18$)	Kruskal- Wallis Test	Post-hoc tests (Mann-Whitney U Test)
	Mean (SD)	Mean (SD)	Mean (SD)	p	
Remember	1.8 (1.94)	1.58 (2.40)	0.44 (0.83)	.193	
Understand	4.6 (3.01)	3.25 (1.79)	1.33 (1.29)	.005**	2>1** 3>1*
Apply	6.2 (4.62)	4.08(3.20)	2.5(1.17)	.177	
analyze	8.4 (3.2)	5.92(1.66)	3.89(2.33)	.006**	2>1* 3>1**
Evaluate	44.2 (13.32)	8.83(5.15)	1(1.20)	.000***	2>1*** 3>1*** 3>2***
Create	3 (0.63)	2.75(1.69)	1.83(1.12)	.094	3>1*
Off-topic	15.8 (10.93)	7(4.20)	2.72(2.88)	.002**	2>1** 3>1**

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

The proportion distributions of the three categories of students are presented in Figure 5. Core students experienced the richest cognitive behaviors and the highest percentage of higher-order cognitive behaviors (66.19%) compared with the other two categories of students (semi-peripheral students: 52.37%; peripheral students: 48.99%). Peripheral students experienced the highest percentage of lower-order cognitive behaviors (31.17%) compared with the other two categories of students (core students: 15%; semi-peripheral students: 26.68%). Semi-peripheral students experienced the highest percentage of *off-topic* behaviors (20.95%) compared with the other two categories of students (core students: 18.81%; peripheral students: 19.84%).

Kruskal–Wallis tests were conducted to compare the three categories of students in terms of each type of cognitive behavior. As presented in Table 2, four categories, *understand*, *analyze*, *evaluate*, and *off-topic*, are significantly different ($p < .01$, $p < .01$, $p < .001$, and $p < .01$, respectively).

Figure 5. The proportion distributions of cognitive behaviors in different students

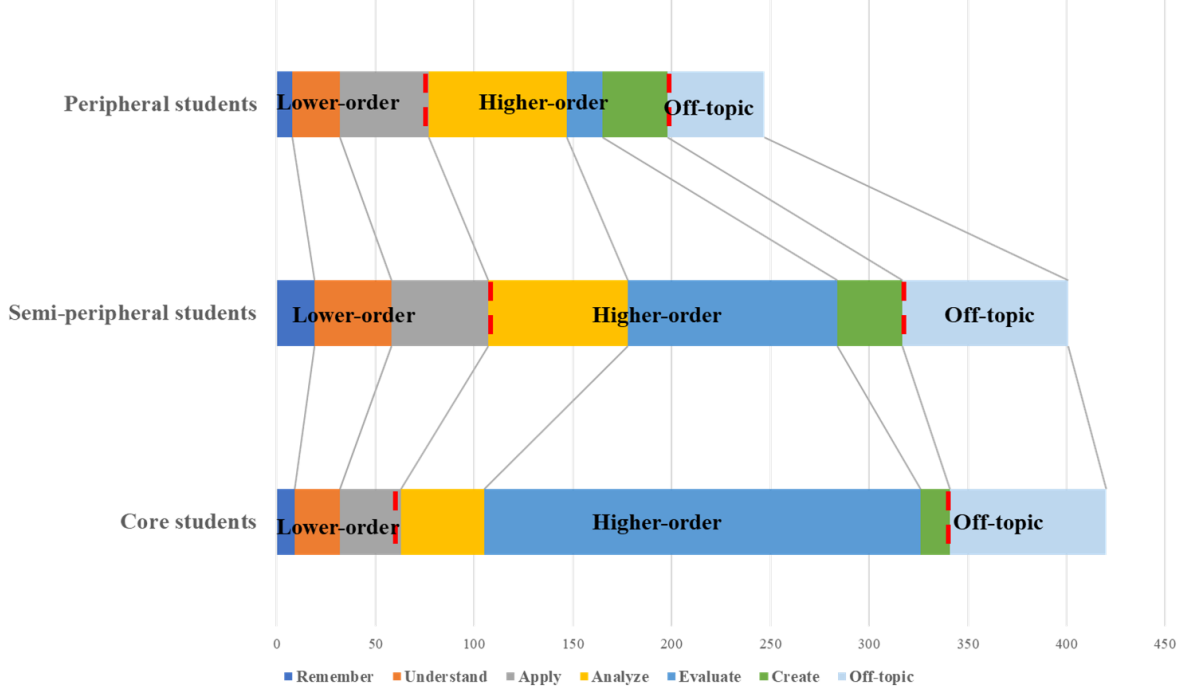
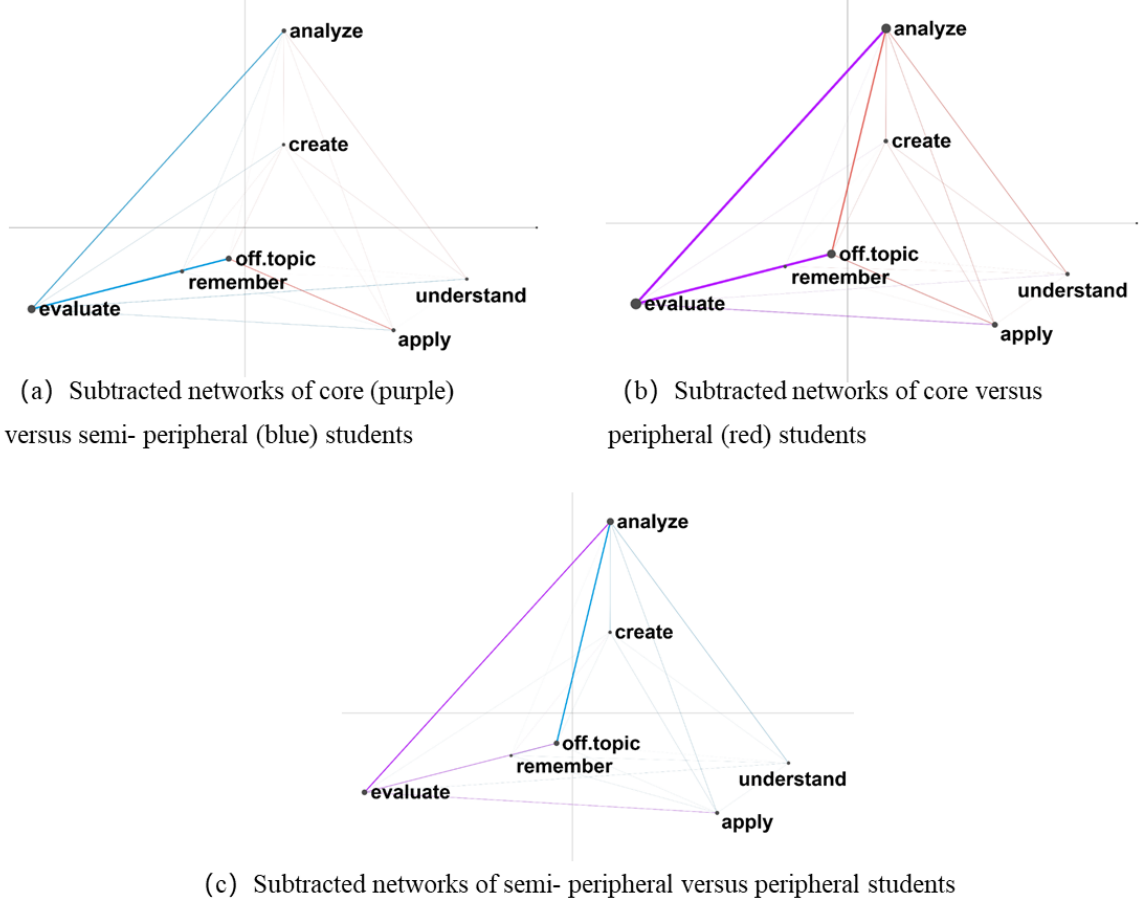


Figure 6. Subtracted networks of different categories of students



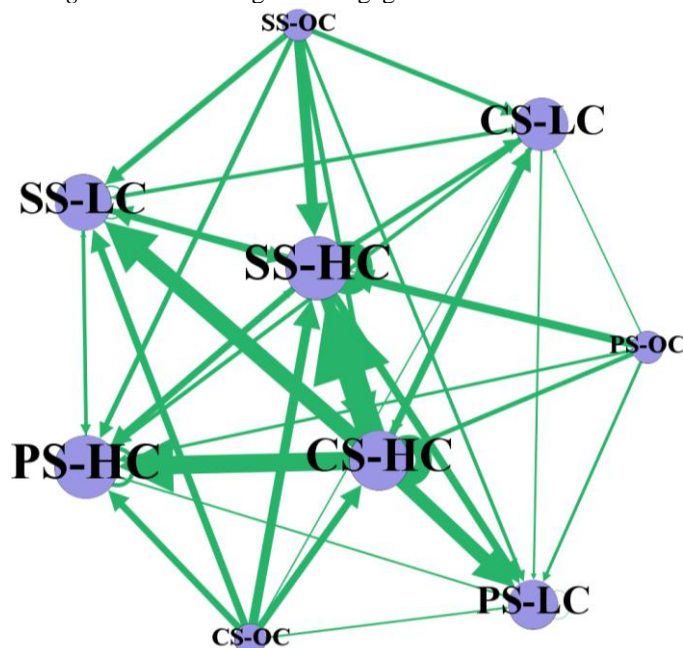
Furthermore, pairwise Mann–Whitney U Tests showed that statistical significance existed for cognitive behavior. Although *remember*, *apply*, and *create* among the three categories of students did not show significant differences (*remember*: $\chi^2(2, N = 35) = 3.295, p = .193$; *apply*: $\chi^2(2, N = 35) = 3.461, p = .177$; *create*: $\chi^2(2, N = 35) = 4.736, p = .094$), Mann–Whitney U Tests revealed a significant difference between the core and peripheral students in terms of *create* ($U = 18, z = -2.091, p < .05$).

To explore the differences in cognitive behavior across the three types of students, a series of ENA were conducted. Figure 6 shows the subtracted networks of three types of students among different cognitive behaviors. Lower-order and *off-topic* cognitive behaviors were mainly in quadrants III and IV, and higher-order cognitive behaviors were mainly in quadrants I and III. Referring to Figure 6, compared to the other two types of students, core students had stronger connections between *analyze* and *evaluate* and between *evaluate* and *off-topic*. In contrast, significant associations were uncovered between *off-topic* and *apply* in the network for the peripheral students. In terms of the connection between *analyze* and *off-topic*, although the subtracted networks of semi-peripheral and peripheral students showed no difference, both of them were stronger than those of core students.

4.2. Interactive pattern of students' social-cognitive engagement in online discussions

To answer RQ2, we constructed social-cognitive engagement from two perspectives—social network position and cognitive level—to further describe the interactive pattern of social-cognitive engagement and to explore the relationship between social-cognitive engagement and academic performance.

Figure 7. Social-cognitive engagement interactive network



Note. Node size represents degree. The directed lines between nodes represent the frequency and direction of interactions.

After constructing social-cognitive engagement, we drew the interactive network diagram of social-cognitive engagement. Figure 7 shows a social-cognitive engagement interactive network that reveals the interaction details of the various social-cognitive engagement. It is clear that higher-order behaviors accounted for the most in terms of number and type of interactions, whereas *off-topic* cognitive behaviors involved fewer interactions. For the same level of cognitive behavior, different categories of students had all kinds of interaction characteristics. Specifically, for higher-order cognitive behavior, core students (CS-HC) were primarily involved in the response process, especially responding to semi-peripheral students. Semi-peripheral students (SS-HC) both responded to others and received responses from others. Peripheral students (PS-HC) mainly received responses from others. In addition, core and semi-peripheral students exhibited more self-interaction behaviors. For lower-order cognitive behavior, all students mainly received responses, but a small number of students in each category (CS-LC, SS-LC, and PS-LC) actively replied to others. According to the direction of interaction, all *off-topic* cognitive behaviors of students (CS-OC, PS-OC, and SS-OC) were only involved in the process of responding. In other words, students did not actively interact with content that was not related to the course.

To measure the importance of different social-cognitive engagements in interactive networks, PageRank, hub, and authority were calculated, as presented in Table 3. Regardless of the category of students, *off-topic* cognitive behaviors (CS-OC, SS-OC, and PS-OC) had the smallest PageRank value, the largest hub value, and the smallest of authority value, implying that the content reflecting *off-topic* cognitive behavior is the least important in the

interaction. Higher-order cognitive behaviors (PS-HC and SS-HC) had the largest PageRank value and the largest authority value, which means that higher-order cognitive behaviors were the most important in the interaction.

Table 3. The network importance of different social-cognitive engagement

Social-cognitive engagements	PageRank	Authority	Hub
CS-OC	0.016667	0	0.397161
PS-OC	0.016667	0	0.397161
SS-OC	0.016667	0	0.397161
PS-LC	0.111358	0.407933	0.136098
CS-LC	0.111358	0.396914	0.332789
SS-LC	0.111358	0.407933	0.264842
CS-HC	0.114595	0.396914	0.332789
PS-HC	0.146969	0.419583	0.329112
SS-HC	0.146969	0.419583	0.329112

Kruskal–Wallis tests showed that social-cognitive engagements were statistically significant in terms of academic performance ($p < .001$). Pairwise Mann–Whitney U tests were performed as post-hoc analyses to find specific differences. A total of 36 comparisons were made, and a total of 27 data pairs with significant differences were found. Table 4 shows the results of data pairs with their differences. On the one hand, more achievement differences existed between social-cognitive engagement involving core and peripheral students (e.g., PS-OC, PS-LC, CS-OC, and CS-HC) and other social-cognitive engagement. On the other hand, social-cognitive engagement involving *off-topic* (e.g., PS-OC and CS-OC) differed more in terms of academic performance than did other social-cognitive engagements. Social-cognitive engagement related to semi-peripheral students did not differ from one to another related to semi-peripheral students in terms of academic performance.

Table 4. The results of data pairs related to social-cognitive engagement with differences

Sample 1	Sample 2	U	Z	<i>p</i>	Sample 1	Sample 2	U	Z	<i>p</i>
PS-OC	PS-LC	1017.5	-2.660	.008**	PS-HC	SS-LC	1556	-9.604	.000***
PS-OC	SS-OC	624	-5.411	.000***	PS-HC	SS-HC	4210	-10.955	.000***
PS-OC	SS-LC	760	-5.857	.000***	PS-HC	CS-OC	264	-11.447	.000***
PS-OC	SS-HC	2014	-5.75	.000***	PS-HC	CS-LC	880	-8.601	.000***
PS-OC	CS-OC	144	-8.406	.000***	PS-HC	CS-HC	2772	-14.420	.000***
PS-OC	CS-LC	480	-5.418	.000***	SS-OC	CS-OC	1519	-4.514	.000***
PS-OC	CS-HC	1512	-8.255	.000***	SS-OC	CS-HC	7397	-3.621	.000***
PS-LC	SS-OC	385	-8.446	.000***	SS-LC	CS-OC	2317	-3.58	.000***
PS-LC	SS-LC	482	-8.99	.000***	SS-LC	CS-HC	9912	-3.502	.000***
PS-LC	SS-HC	1294	-9.885	.000***	SS-OC	CS-OC	3506	-6.957	.000***
PS-LC	CS-OC	102	-9.886	.000***	SS-OC	CS-HC	18631	-6.905	.000***
PS-LC	CS-LC	340	-7.913	.000***	CS-OC	CS-LC	1512	-3.108	.000***
PS-LC	CS-HC	1071	-11.386	.000***	CS-LC	CS-HC	6608	-2.538	.011*
PS-HC	SS-OC	1324	-8.609	.000***					

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

4.3. Evolution of students' social-cognitive engagement during different phases of online discussions

To answer RQ3, we employed ENA to uncover the evolution of social-cognitive engagement during online discussions.

ENA characterized all the chronological networks so that the evolution of social-cognitive engagement during different phases could be compared visually and statistically. Figure 8 is the social-cognitive engagement networks for the six phases. The number of posts published in the six phases was roughly approximate. Figure 8 shows that social-cognitive engagements related to core students were mainly found in quadrant IV and that social-cognitive engagements related to semi-peripheral students were mainly found in quadrant I. Quadrants II and quadrants III were scattered with social-cognitive engagements related to peripheral students. Broadly, social-cognitive engagements related to peripheral and non-peripheral students were distinguished by the Y-axis, whereas the X-axis distinguished between social-cognitive engagements related to core students and that to semi-peripheral students. Table 5 shows the coordinates of the centroids of the six phases. The centroid takes into

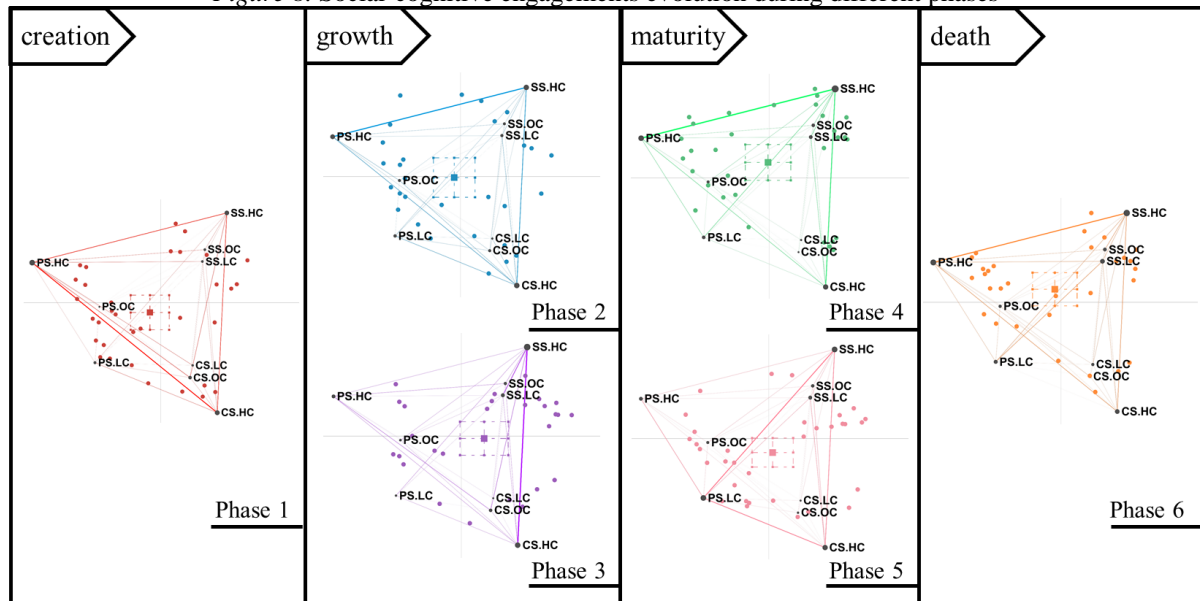
account the weights of the connections between cognitive elements and can be represented as a corresponding plotted point (Bressler et al., 2019). The coordinates of phases 1, 2, 4, and 6 were on the negative axis of the X-axis, which indicated that there was a relatively strong co-occurrence between the social-cognitive engagement related to peripheral students in these phases. Conversely, the coordinates of phases 1, 2, 3, and 5 were on the negative axis of the Y-axis, which indicated that there was a relatively strong connection between the social-cognitive engagement related to core students in these phases. For instance, the position of the phase 1 centroid was shown on the Cartesian coordinate system as (-0.1, -0.09). Correspondingly, there was a greater co-occurrence of social-cognitive engagement related to core students and peripheral students in phase 1.

Table 5. The coordinates of the six phases' centroids

Phases	Dimension	
	X	Y
1	-0.1	-0.09
2	-0.06	-0.01
3	0.22	-0.02
4	-0.01	0.15
5	0.04	-0.14
6	-0.06	0.13

As suggested by Iriberry and Leroy (2009), the learning process can be divided into five periods: inception, creation, growth, maturity, and death. Inception involves the design process of the online forum; during this period, the students have not yet entered the forum to participate in the interaction and therefore were not considered in this study. In Figure 8, the six phases were summarized in four periods. The changes in the co-occurrence of social-cognitive engagements indicated a learning diagram over the four periods. The findings revealed that there was a joint connection between high-order cognitive behaviors at all phases (e.g., co-occurrences between SS-HC and CS-HC).

Figure 8. Social-cognitive engagements evolution during different phases



A two-sample t-test was used to determine whether there were significant differences in the position of the phase centroid between two adjacent phases. The results indicated that phase 5 was significantly different from phases 4 ($t = -2.61, p = .01$) and 6 ($t = -2.50, p = .02$) on the Y-axis. Therefore, the development of the social-cognitive engagement network over the six phases did not follow a straight upward route but rather followed a nonlinear route: both the cognitive level and the participation of different types of students reached a relatively high level but returned to intermediate levels at phase 6, the last period, referred to death.

In the creation period, the connection between CS-HC and PS-HC (connection coefficient: 0.22) was the strongest, while some of the other connections focused on CS-HC and SS-HC (connection coefficient: 0.16), and PS-HC and SS-HC (connection coefficient: 0.17). Stronger occurrence relationships were found in the growth period for the connection between PS-HC and SS-HC (connection coefficient: 0.25) in phase 2, and CS-HC and SS-HC (connection coefficient: 0.27) in phase 3. The strongest co-occurrence of SS-HC and PS-HC (connection

coefficient: 0.30) occurred in phase 4. As the death period, the distinctive co-occurrence relationships in phase 6 were PS-HC and SS-HC (connection coefficient: 0.24). As Iriberry and Leroy (2009) pointed out, discussion forums would experience poor member participation, insufficient quality of content, and weak ties between members. The fewer number of co-occurrence relationships in phase 6 was a reflection of the death.

5. Discussion

The results captured the dynamic complex interaction process in the forum at a fine-grained level that combines social and cognitive perspectives. Regarding the three research questions that range from the relationship between social and cognitive aspects to joint modeling, a detailed discussion is as follows.

5.1. Relationship between social and cognitive aspects in online discussions

With regard to the relationship between social and cognitive aspects, the results of the current study indicate that students' social network position is a vital indicator for the contributions to knowledge construction. It is clear that core students made more contributions to the overall cognitive discussion. The greatest contributions on *off-topic* were made by semi-peripheral students, and peripheral students made the greatest contribution to lower-order cognitive behaviors. Peripheral students only posted their ideas and rarely responded to the comments of others, so their contribution to the development of group knowledge construction and interactive networks was limited. The results are similar to those of previous studies. For example, regarding 20 students as the study subjects, Ouyang and Chang (2019) examined the relationships between social participatory roles and cognitive engagement levels. Results indicated that peripheral students had the lowest average scores on cognitive engagement levels. Although peripheral students were at the periphery of the social network and rarely responded to peer comments in this study, this did not mean that they rarely received comments. In this study, S26, who was a peripheral student, never responded to other students but received responses from three core students: S2, S12, and S13. Each peripheral student received a response from core students or some semi-peripheral students. The results revealed similar findings in previous studies that core students should also be reasonably well connected to peripheral students (Rombach et al., 2014). Overall, compared to other students, core students in social networks made more contributions to knowledge construction.

5.2. Interactive pattern of students' social-cognitive engagement in online discussions

With regard to the interactive pattern of social-cognitive engagement in an online discussion, the results indicated that different cognitive levels made different contributions to interaction. The different cognitive levels manifested by students in different network positions showed various interactive characteristics within the overall interactive network, but from an overall perspective, higher-order cognitive behaviors were more likely to trigger positive and more interactions than other cognitive behaviors, even the higher-order cognitive behaviors manifested by peripheral students. Take S7 as an example, S7 published a total of 14 discussion posts, 8 of which were coded as higher-order cognitive. These 8 posts received a total of 9 replies, while the other 6 non-higher-order cognitive posts did not receive replies. Consequently, in addition to the possibility that a climate of knowledge sharing and group cohesion could be formed with the help of social interaction, the development of knowledge construction also, in turn, influenced social interactions. The frequent interaction of higher-order cognitive behaviors may originate from the reflective and permanent character of online discussions. Students synthesize ideas and integrate them with their existing knowledge through continuous reflection. Precise cognitive behaviors that contain reflective meaning are the conditions for the persistence of interaction.

In addition, based on the social interactive network of these nine types of social-cognitive engagement and the calculation of their importance, social network position also affects the interactive characteristics. The social-cognitive engagement: CS-LC is a typical representative that did not exhibit self-interaction behavior. This phenomenon complements the "rich club" (Vaquero & Cebrian, 2013). Active students engage in the forum not only to build rich peer connections through persistent interactions but also to selectively respond to posts that made deep cognitive contributions. In the current study, although core students will initiate interactive connections with peripheral students, core students prefer to respond to higher-order cognitive posts made by peripheral students rather than lower-order cognitive posts and off-topic posts. Overall, different social-cognitive engagements showed different interactive characteristics, which were influenced by both social network position and cognitive level.

5.3. Evolution of students' social-cognitive engagement during different phases of online discussions

In terms of the evolution of social-cognitive engagement in online discussions, the results demonstrated nonlinear development over time. A significant higher-order cognitive tendency (e.g., CS-HC) was observed in the creation and growth periods of the online forum. This means that, when core students post higher-order cognitive posts, students' social engagement behaviors in forums will be effectively promoted, thereby improving the social network cohesion of forums. The connections among different social-cognitive engagements changed significantly with the increase in discussion activities over time, especially in phase 5. As suggested by Iriberri and Leroy (2009), a critical mass of members and member-generated content is reached at maturity, and teachers provide teaching interventions such as rewarding members or managing subgroups. At death, the forum experienced poor participation and unorganized contributions, and termination of interaction may be eminent. Discussion tasks and topics could have an impact on students' cognition, and the nonlinear development pattern could be caused by the design of topics. Overall, the four periods were characterized by the specific characterizations of the network structure of social-cognitive engagement changes in the current study.

6. Conclusions, limitations, and future research

The main contribution of this study is that the social-cognitive engagement jointly characterizes students' social and cognitive aspects, allowing us to gain a deeper understanding of knowledge construction from the perspectives of social and cognitive connections and their joint impact on online forums. The construction process provides meaningful insights for joint analysis in subsequent research. By combining SNA, CA, and ENA, the results showed that students' social network position was a vital indicator of their contributions to knowledge construction, especially core students who contributed more to knowledge construction. In addition, higher-order cognitive behaviors made more contributions to interaction. In summary, the interactive characteristics of social-cognitive engagement were affected by both social network position and cognitive level. Apart from this, the nonlinear trajectory of social-cognitive engagement uncovered its evolutionary trend. Significant changes in the connections between different social-cognitive engagements can indicate the dynamic evolution of the forum. Maturity is more informative compared to other periods, that is, there were significant differences in the position of the centroid between maturity and adjacent periods. Compared to other periods, the connection coefficient of social-cognitive engagements at maturity was not high and there was no significant co-occurrence characteristic. Based on the results, this study provides methodological implications for multi-perspective analysis and practical suggestions for teachers to improve students' social and cognitive levels.

Our study provides researchers with methodological implications from multi-perspective analysis of online forums. The combination of multiple methods, especially SNA and ENA, is a useful method to get a more comprehensive view of student engagement characteristics. Social-cognitive engagement jointly characterizes the social and cognitive aspects of students, rather than describing learning characteristics from a single social or cognitive perspective. Students were classified as core, semi-peripheral, and peripheral students according to coreness from a social perspective, and posts were defined as higher-order, lower-order, and *off-topic* cognitive behaviors from a cognitive perspective. This is an idea of joint analysis that can even be extended to other perspectives, such as sentiment in the text. Research examining sentiment evolution with different interactions is scant (Huang et al., 2021), and this study may inspire a combined cognitive, social, and emotional analysis.

In addition, our study provides practical suggestions for teachers to strengthen the development of interactive online discussions and increase students' cognitive levels. Students may provide higher-order cognitive behavior even if they are at the periphery of the social network. When students see themselves as creators of discussion-based learning, they are more inclined to actively participate in the learning process. Teachers could encourage students to participate in top-level planning, decision-making, and learning coordination activities (Ouyang & Chang, 2019) to bring peripheral students closer to the core part of the social network. In addition, to improve the quality of student interaction, teachers can provide scaffolding tools (Lin et al., 2020) and relinquish control of the forum as appropriate (Ouyang & Scharber, 2017) during the learning process. Reasonable instructional designs can be provided based on the evolutionary patterns of the network, such as rewarding mechanisms at maturity.

This study has some limitations that should be noted. First, although there are similar samples in previous studies, for example, Huang et al. (2021) used texture data of 38 students and Ouyang and Chang (2019) used discourse data of 20 learners, the generalizability of the results might be limited due to the sample data. Thus, larger samples should be incorporated in future studies to make the results more representative. Second, there may be some face-to-face interactions that might influence students' online interactions during the 14-week

online discussion activities. Future research that integrates online and non-online interactions to obtain a full understanding of students' interactive behaviors is necessary. Moreover, this study constructed social-cognitive engagement based on students' network position, and other social network properties, such as social roles, can also be used. It would be of great significance to investigate students' social roles and their relationship with learning achievements in future research.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62107016, 62077017, 61977030, 61937001), the Research funds from the Humanities and Social Sciences Foundation of the Ministry of Education (Grant No. 21YJC880057), Hubei Provincial Natural Science Foundation of China (Grant No. 2021CFB140), and the Financially supported by self-determined research funds of CCNU from the colleges' basic research and operation of MOE Fundamental Research Funds of the Central Universities (Grant Nos. CCNU20TS032, 30106200548, CCNU21XJ034). We would like to thank AJE (<https://www.aje.com/>) for English language editing.

References

- Bereiter, C. (2002a). *Education and mind in the knowledge age*. Lawrence Erlbaum Associates.
- Bereiter, C. (2002b). Liberal education in a knowledge society. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp.11–33). Open Court.
- Borgatti, S. P., & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, 21(4), 375–395. [https://doi.org/10.1016/S0378-8733\(99\)00019-2](https://doi.org/10.1016/S0378-8733(99)00019-2)
- Bressler, D. M., Bodzin, A. M., Eagan, B., & Tabatabai, S. (2019). Using epistemic network analysis to examine discourse and scientific practice during a collaborative game. *Journal of Science Education and Technology*, 28(5), 553–566. <https://doi.org/10.1007/s10956-019-09786-8>
- Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, 116, 93–109. <https://doi.org/10.1016/j.compedu.2017.08.007>
- Fougt, S. S., Siebert-Evenstone, A., Eagan, B., Tabatabai, S., & Misfeldt, M. (2018). Epistemic network analysis of students' longer written assignments as formative/summative evaluation. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 126–130. <https://doi.org/10.1145/3170358.3170414>
- Garrison, D. R., Anderson, T., & Archer, W. (2010). The First decade of the community of inquiry framework: A Retrospective. *The Internet and Higher Education*, 13(1), 5–9. <https://doi.org/10.1016/j.iheduc.2009.10.003>
- Gašević, D., Joksimović, S., Eagan, B. R., & Shaffer, D. W. (2019). SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, 92, 562–577. <https://doi.org/10.1016/j.chb.2018.07.003>
- Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. 37–56). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7_2
- Hong, H. Y., Chen, F. C., Chai, C. S., & Chan, W. C. (2011). Teacher-education students' views about knowledge building theory and practice. *Instructional Science*, 39(4), 467–482. <https://doi.org/10.1007/s11251-010-9143-4>
- Huang, C., Han, Z., Li, M., Wang, X., & Zhao, W. (2021). Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis. *Australasian Journal of Educational Technology*, 37(2), 81–95. <https://doi.org/10.14742/ajet.6749>
- Iriberry, A., & Leroy, G. (2009). A Life-cycle perspective on online community success. *ACM Computing Surveys*, 41(2), 11:1–11:29. <https://doi.org/10.1145/1459352.1459356>
- Ke, F., & Xie, K. (2009). Toward deep learning for adult students in online courses. *The Internet and Higher Education*, 12(3), 136–145. <https://doi.org/10.1016/j.iheduc.2009.08.001>
- Lin, K. Y., Hong, H.-Y., & Chai, C. S. (2014). Development and validation of the knowledge-building environment scale. *Learning and Individual Differences*, 30, 124–132. <https://doi.org/10.1016/j.lindif.2013.10.018>

- Lin, P.-C., Hou, H.-T., & Chang, K.-E. (2020). The Development of a collaborative problem solving environment that integrates a scaffolding mind tool and simulation-based learning: An Analysis of learners' performance and their cognitive process in discussion. *Interactive Learning Environments*, 0(0), 1–18. <https://doi.org/10.1080/10494820.2020.1719163>
- Liu, C. H., & Matthews, R. (2005). Vygotsky's philosophy: Constructivism and its criticisms examined. *International Education Journal*, 6(3), 386–399.
- Ouyang, F., & Chang, Y.-H. (2019). The Relationships between social participatory roles and cognitive engagement levels in online discussions. *British Journal of Educational Technology*, 50(3), 1396–1414. <https://doi.org/10.1111/bjet.12647>
- Ouyang, F., & Scharber, C. (2017). The Influences of an experienced instructor's discussion design and facilitation on an online learning community development: A Social network analysis study. *The Internet and Higher Education*, 35, 34–47. <https://doi.org/10.1016/j.iheduc.2017.07.002>
- Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of innovative knowledge communities and three metaphors of learning. *Review of Educational Research*, 74(4), 557–576. <https://doi.org/10.3102/00346543074004557>
- Peng, X., Han, C., Ouyang, F., & Liu, Z. (2020). Topic tracking model for analyzing student-generated posts in SPOC discussion forums. *International Journal of Educational Technology in Higher Education*, 17(1), 35. <https://doi.org/10.1186/s41239-020-00211-4>
- Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education*, 143, 103673. <https://doi.org/10.1016/j.compedu.2019.103673>
- Popescu, E., & Badea, G. (2020). Exploring a community of inquiry supported by a social media-based learning environment. *Educational Technology & Society*, 23(2), 61–76.
- Rombach, M. P., Porter, M. A., Fowler, J. H., & Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 74(1), 167–190. <https://doi.org/10.1137/120881683>
- Sfard, A. (1998). On Two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2), 4–13. <https://doi.org/10.2307/1176193>
- Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A Tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3), 9–45. <https://doi.org/10.18608/jla.2016.33.3>
- Swiecki, Z., & Shaffer, D. W. (2020). iSENS: An Integrated approach to combining epistemic and social network analyses. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 305–313). <https://doi.org/10.1145/3375462.3375505>
- Tan, S. C., Wang, X., & Li, L. (2022). The Development trajectory of shared epistemic agency in online collaborative learning: A Study combining network analysis and sequential analysis. *Journal of Educational Computing Research*, 59(8), 1655–1681. <https://doi.org/10.1177/07356331211001562>
- Teo, H. J., Johri, A., & Lohani, V. (2017). Analytics and patterns of knowledge creation: Experts at work in an online engineering community. *Computers & Education*, 112, 18–36. <https://doi.org/10.1016/j.compedu.2017.04.011>
- Tirado, R., Hernando, Á., & Aguaded, J. I. (2015). The Effect of centralization and cohesion on the social construction of knowledge in discussion forums. *Interactive Learning Environments*, 23(3), 293–316. <https://doi.org/10.1080/10494820.2012.745437>
- Vaquero, L. M., & Cebrian, M. (2013). The Rich club phenomenon in the classroom. *Scientific Reports*, 3(1), 1174. <https://doi.org/10.1038/srep01174>
- Wang, K., Chen, L., & Anderson, T. (2014). A Framework for interaction and cognitive engagement in connectivist learning contexts. *International Review of Research in Open and Distance Learning*, 15, 121–141. <https://doi.org/10.19173/irrodl.v15i2.1709>
- Yücel, Ü. A., & Usluel, Y. K. (2016). Knowledge building and the quantity, content and quality of the interaction and participation of students in an online collaborative learning environment. *Computers & Education*, 97, 31–48. <https://doi.org/10.1016/j.compedu.2016.02.015>
- Zhang, J., Skryabin, M., & Song, X. (2016). Understanding the dynamics of MOOC discussion forums with simulation investigation for empirical network analysis (SIENA). *Distance Education*, 37(3), 270–286. <https://doi.org/10.1080/01587919.2016.1226230>
- Zhang, S., Liu, Q., Chen, W., Wang, Q., & Huang, Z. (2017). Interactive networks and social knowledge construction behavioral patterns in primary school teachers' online collaborative learning activities. *Computers & Education*, 104, 1–17. <https://doi.org/10.1016/j.compedu.2016.10.011>

Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness

Siu-Cheung Kong^{1,2*}, William Man-Yin Cheung² and Guo Zhang²

¹Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR // ²Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong SAR // siucheungkong@gmail.com // williamcheung@eduhk.hk // gzhang@friends.eduhk.hk

*Corresponding author

(Submitted November 29, 2021; Revised March 26, 2022; Accepted April 25, 2022)

ABSTRACT: Emerging research is highlighting the importance of fostering artificial intelligence (AI) literacy among educated citizens of diverse academic backgrounds. However, what to include in such literacy programmes and how to teach literacy is still under-explored. To fill this gap, this study designed and evaluated an AI literacy programme based on a multi-dimensional conceptual framework, which developed participants' conceptual understanding, literacy, empowerment and ethical awareness. It emphasised conceptual building, highlighted project work in application development and initiated teaching ethics through application development. Thirty-six university students with diverse academic backgrounds joined and completed this programme, which included 7 hours on machine learning, 9 hours on deep learning and 14 hours on application development. Together with the project work, the results of the tests, surveys and reflective writings completed before and after these courses indicate that the programme successfully enhanced participants' conceptual understanding, literacy, empowerment and ethical awareness. The programme will be extended to include more participants, such as senior secondary school students and the general public. This study initiates a pathway to lower the barrier to entry for AI literacy and addresses a public need. It can guide and inspire future empirical and design research on fostering AI literacy among educated citizens of diverse backgrounds.

Keywords: Application development, Artificial intelligence literacy, Conceptual framework, Ethical awareness, University students

1. Introduction

Fostering artificial intelligence (AI) literacy for all citizens has become increasingly crucial, given AI's potential to reshape the competitive landscape and its relevance to individuals' lives and work (Fosso Wamba et al., 2021; JRC & OECD, 2021; WIPO, 2019). However, few studies have comprehensively examined how and what exactly to teach to educate citizens of diverse backgrounds.

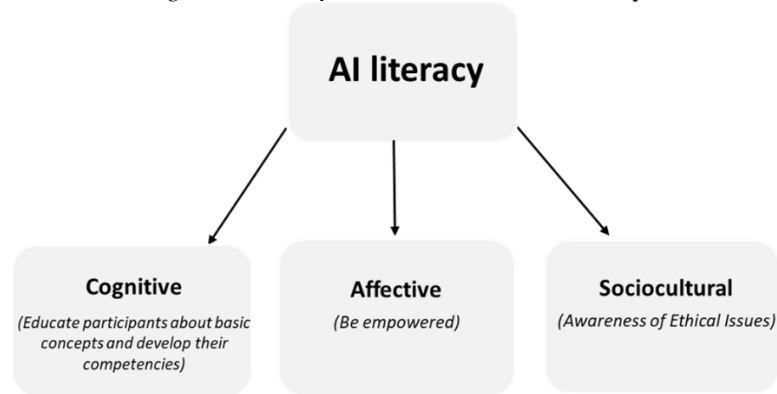
Most studies of conceptual teaching involve mathematical formulae and programming codes, focusing primarily on computer science majors and students with programming knowledge (Green, 2021; Pouly et al., 2019; Stadelmann et al., 2021; Tedre et al., 2021). This approach creates a barrier to literacy amongst the public (Long & Magerko, 2020). While ethical issues related to AI have received increased attention (Ashok et al., 2022; Jobin et al., 2019; Kuipers, 2020; Mehrabi et al., 2021; Prunkl, 2022), ethics thus far have rarely been an explicit component of AI courses (Saltz et al., 2019), and limited information is available on the ethical considerations covered in AI classes (Garrett et al., 2020).

To fill this research gap and to serve social equity and sustainable development goals (Kong et al., 2021b; OECD, 2018a; Vinuesa et al., 2020), this study develops an AI programme that focuses on conceptual understanding, literacy, empowerment and ethical awareness. The literacy development framework presented here focuses on conceptual building, emphasising project work in application development and enhancing participants' awareness of the ethical considerations arising from such work. This study reports the process of designing, implementing and evaluating this AI literacy programme.

2. Background

We follow the conceptual framework of AI literacy from Kong and Zhang (2021) (see Figure 1). This framework is comprised of three dimensions: the cognitive dimension; the affective dimension; and the sociocultural dimension.

Figure 1. Conceptual framework of AI literacy



The cognitive dimension involves teaching major fundamental AI concepts, particularly machine learning and deep learning, and how to use them to evaluate and understand the real world. These concepts have profound societal impacts and are essential to fostering AI literacy (OECD, 2018b; Touretzky et al., 2019; Wong et al., 2020). By understanding these concepts, learners should be able to evaluate AI artefacts in their lives and the impacts of the technology, then apply the concepts to understand the AI-permeated world, and form their attitudes and responses accordingly.

The affective dimension serves to empower participants so they can participate with confidence in the digital world. It contains four components: grasping the value of AI (Thomas & Velthouse, 1990); perceiving the social impact of AI (Frymier et al., 1996); believing in one's ability to produce novel AI ideas and solutions (Paulus & Brown, 2003); and being confident in one's competence in engaging with AI (Bandura, 1982). This four-factor model (meaningfulness, impact, creative self-efficacy and AI self-efficacy) is consistent with the idea of future literacy from the United Nations Educational, Scientific and Cultural Organization (UNESCO), which aims to strengthen learners' imagination and prepare them for change (Yi, 2021). Our initiative aims to develop participants' self-confidence in conducting AI-related activities, educate them about AI's significance and societal impacts, and enhance their digital creativity.

Finally, the sociocultural dimension concerns the ethical use of AI. Our course followed the ethical principles outlined in Kong and Zhang (2021), which was built on those stated in the Belmont Report (NCPHS, 1978): (1) the use of AI should not violate human autonomy; (2) AI's benefits should outweigh its risks; and (3) AI's benefits and risks should be distributed equally. These three principles (autonomy, beneficence/non-maleficence and fairness) have also been covered by recent AI ethical frameworks (Floridi & Cowls, 2019; HLEG, 2019; OECD, 2019). As effective guidelines to follow, they serve as the constructs of the ethical consideration survey detailed in Section 3.4 below.

This multidimensional conceptual framework informs our design, development and evaluation of this literacy programme. Using this framework, this study focused on the three research questions: (1) Can the AI literacy programme address AI concepts and literacy? (2) Will participants feel empowered after completing the AI literacy programme? and (3) Can the AI literacy programme foster participants' ethical awareness?

3. Methodology

3.1. Course participants

We launched a literacy programme at a Hong Kong university for convenience sampling. A total of 36 university students from diverse backgrounds joined the programme. Twenty-three were female and thirteen were male. Seventy-five per cent of the participants were enrolled in bachelor's degree programmes, including students in their first, second, third and fourth years of study. The remaining participants were from postgraduate or higher diploma programmes. As shown in Table 1, the participants came from a wide range of academic backgrounds, namely Mathematics, Information and Communication Technology, Health Education, Chinese Language Studies, Psychology, the Sciences (Natural Science & STEM Education), English Language Studies, General Studies, Music, History, Global and Environmental Studies and Global and Hong Kong Studies.

Table 1. Distribution of programme participants' academic backgrounds

Academic background	Number (percentage)	Academic background	Number (percentage)
Mathematics	8 (22.22%)	English Language Studies	2 (5.56%)
Information and Communication Technology	5 (13.89%)	General Studies	2 (5.56%)
Health Education	4 (11.11%)	Music	2 (5.56%)
Chinese Language Studies	4 (11.11%)	History	1 (2.78%)
Psychology	3 (8.33%)	Global and Environmental Studies	1 (2.78%)
The Sciences (Natural Science & STEM Education)	3 (8.33%)	Global and Hong Kong Studies	1 (2.78%)
Total			36 (100%)

3.2. Curriculum

The programme consisted of three courses: Machine Learning, Deep Learning and Developing Artificial Intelligence Applications. The first two courses develop conceptual understanding of two important AI areas (Kong & Zhang, 2021; Kong et al., 2021b), thus fostering AI literacy in the cognitive dimension. The third course further develops AI literacy through applying acquired concepts to project work. This project work in turn serves as a concrete example to reflect on ethical issues, thus covering the sociocultural dimension. The affective domain is also enhanced as participants can feel more empowered with more understanding of AI throughout all three courses.

3.2.1. Course 1: Machine learning

Course 1 introduced the concepts and some related algorithms in both supervised and unsupervised learning. An overview of AI's development was first provided, followed by concepts of strong and weak AI. The participants were encouraged to voice their thoughts on AI's impact on society.

With this foundation, the participants then discussed the "five steps of machine learning," together with hands-on experience using these steps to perform image recognition on an online platform. Afterwards, the participants learned about two instances of supervised learning, "regression" and "classification," through examples and hands-on experience. Finally, this course covered the concept and working principles of unsupervised learning by applying k-means clustering in a series of case studies (Kong et al., 2021b).

In teaching these concepts, we emphasised conceptual building from the beginning: we used analogies and real-life scenarios rather than programme codes and mathematical formulae to foster students' conceptual understanding (Kong et al., 2021b). This allowed the course participants to understand the fundamental concepts of AI and the rationale that underlie them, thereby simplifying the learning process while deepening their conceptual understanding.

3.2.2. Course 2: Deep learning

In the same vein, Course 2 developed the participants' conceptual understanding of deep learning. The course covered several topics, including data cleaning, data augmentation, neural networks, computer vision, deep learning and convolution neural networks. Through reviewing the application of the five steps of machine learning in case studies, the course presented the ideas of data cleaning and data augmentation. The concept of neural networks was introduced by explaining the ideas of perception, input layers, hidden layers, output layers and weights, among others. The participants' understanding was deepened through a lab session of training neural networks to learn to distinguish different data points within various data sets. The concept of computer vision was then discussed, as it is commonly applied in neural networks; related applications were shared with the course participants to provide first-hand experience. The participants were also introduced to convolution neural networks through a lab session and various discussions. Finally, the participants were given the opportunity to experience more machine learning tools.

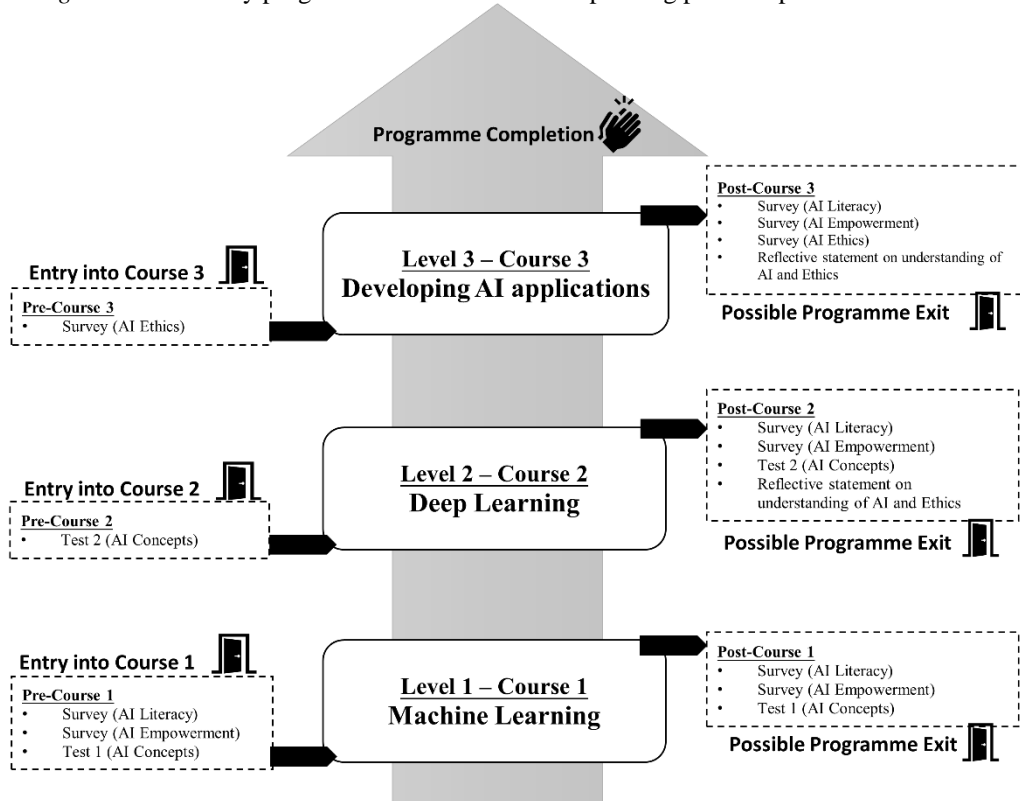
3.2.3. Course 3: Developing AI applications

Course 3 developed the participants' ethical awareness through project work in application development. Before the course, the participants were provided with self-directed reading materials concerning ethical dilemmas and principles of ethics related to AI. Ethical principles on designing and using AI were introduced in the lectures to deepen participants' understanding through real-life examples and discussions. Three additional examples of AI applications were then presented, and the participants discussed the ethical issues involved in these cases in groups. Afterwards, the participants brainstormed their project ideas with preliminary feedback from the course tutors, who later organised individual consultations to provide more detailed instructions on project developments. Sessions introducing different project development platforms and tools, such as Microsoft Azure Machine Learning Studio, Google Teachable Machine and Microsoft Azure QnA Maker, were also offered. With this foundation, the participants began collaborative work on their projects. Each group presented their work in the final session with peer evaluation and discussion on the ethical issues involved in each project, which provided them with more opportunities for reflection, further fostering their ethical awareness. Each group's work was assessed using a rubric that included the discussion of ethical considerations as an important component (see Table 7 and Table 11).

3.3. Course administration

Figure 2 shows the flow of courses in the programme and indicates the corresponding pre- and post-course evaluation activities. These surveys and tests were conducted both before and after the courses to study the participants' progress. Because the content of Course 1 differed from that of Course 2, the AI concepts on tests 1 and 2 were designed according to the relevant content. The participants were asked to write a reflective piece either in English or Chinese on their understanding of AI and related ethical issues. The participants were able to exit the programme at the end of each course.

Figure 2. AI literacy programme courses and corresponding pre- and post-course activities



The number of participants in Courses 1, 2 and 3 were 120, 82 and 36, respectively. In this article, we report the findings related to the 36 participants who completed the whole programme. The number of participants for the three courses decreased as the programme progressed. This could be attributed to the increasing difficulty of each course, leading to the withdrawal of participants less confident with the material. This echoes the rationale that participants who are more empowered are more likely to begin or to continue working on the task at hand and make more effort in AI-related projects (Paulus & Brown, 2003; Kong et al., 2021b).

3.4. Instrument design and use

Quantitative and qualitative data were collected through tests, surveys and reflective writings designed to explore participants' development of AI literacy and to encourage self-reflection on ethical issues. In this study, we present the analyses of participants' responses to the following instruments. (1) AI concepts tests: These tests assessed how an AI literacy programme can develop participants' concepts. (2) AI literacy survey: A survey assessed participants' perceptions of their own AI literacy. (3) AI empowerment survey: This survey evaluated participants' empowerment (self-efficacy, meaningfulness, impact, creativity) after completing the AI literacy programme. (4) Survey on ethical considerations in developing AI applications: This survey was used to assess participants' awareness of ethical issues around AI applications. (5) Focus group interview questions on the AI literacy course: The participants were interviewed about their views on AI literacy. (6) Self-reflections on understanding AI ethics. The participants were asked to write 100 to 200 words in either English or Chinese on their understanding of ethical issues related to AI.

Table 2. Bilingual taxonomy of keywords on AI and ethics used for text mining from course participants' self-reflections

Real-world Examples
dilemmas
dilemma / debate / important questions / 困境 / 辯論 / 重要問題
autonomous car / 自動駕駛汽車
copyrights / rights / careful attention / without their knowledge or consent / remuneration / remunerate / reward / 版權 / 權利 / 知識產權 / 謹慎關注 / 報酬 / 未經他們的知情或同意
decision-making / make decisions / make right decisions / decide / moral decision / choose / right answer / judgement / judge / 作出決定 / 作出正確的決定 / 決定 / 道德決策 / 選擇 / 抉擇 / 正確答案 / 判斷
ethics / ethical consideration / ethical issue / ethical problem / ethical reflection / ethical conundrum / ethical solution / 倫理 / 道德 / 道德考量 / 道德問題 / 道德議題 / 道德反思 / 道德難題 / 符合倫理的解決方案
threats
harm / harmful / risk / risky / safety / safe / safely / data security / consequence / bad / threat / 危害 / 有害 / 風險 / 有風險的 / 安全 / 後果 / 不良 / 威脅
replace / 取代 / 代替
piracy / plagiarism / exploit / personal data / privacy / private information / personal information / 盜版 / 抄襲 / 利用 / 個人數據 / 私隱 / 個人信息 / 個人資料
misuse / misusing / abuse / 濫用
discrimination / discriminatory 歧視
bias / biased / stereotypical representations / prejudice / stereotype / 偏見 / 刻板印象
negative / non-transparent / unexplainable / unjustifiable outcomes / lack of explainability / trouble / problematic / problem / inequity / unfairness / unfair / lack of clarity / 負面 / 不透明 / 無法解釋 / 不合理的結果 / 缺乏可解釋性 / 麻煩 / 問題 / 不公平 / 不夠清晰 / 不可靠
isolation / disintegration / reduction of human-to-human interaction / polarise social relationships / damage the wellbeing of individuals / public welfare / 隔離 / 解體 / 減少人與人之間的互動 / 分化社會關係 / 損害個人福祉 / 公共福利
Principles
guide / principle / framework / legal / 指引 / 原則 / 框架 / 合法的
beneficence / nonmaleficence / benefits should outweigh harm / advantages outweigh disadvantages / 為善 / 毋損害 / 利益應大於傷害 / 優點大於缺點
justice / fairness / fair / equity / 正義 / 公道 / 公義 / 公平 / 合理
accuracy / accurate / reliability / reliable / soundness / sound / good reasons / reasonableness / reasonable / 準確 / 可靠 / 健全 / 充分理由
accountability / autonomy / sustainability / transparency / integrity / accountable / sustainable / transparent / responsible / responsibility / responsibilities / accountable / 問責 / 自主 / 可持續 / 透明 / 完整性 / 負責 / 責任 regulation / regulate / 規管 / 監管 / 規範

For the AI concepts tests, the participants were asked to answer multiple choice questions about AI concepts. The tests were designed and guided by the learning progression set forth in Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001). The AI literacy survey addressed the following themes: "AI concepts," "using AI concepts for evaluation" and "using AI concepts to understand the real world" (Kong & Zhang, 2021). The survey items were designed to evaluate the participants' understanding of concepts and related competencies.

The AI empowerment survey included four components: “AI self-efficacy,” “meaningfulness,” “impact” and “creative self-efficacy” (Kong & Zhang, 2021; Kong et al., 2021b). The ethical consideration survey evaluated the participants’ awareness of ethical issues related to AI applications by employing three components: autonomy, beneficence and fairness (Kong & Zhang, 2021). The survey consisted of 12 questions, with four questions concerning each component. For all of the surveys, the participants were asked to indicate their level of agreement with each item on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). Regarding the reliability of the instruments, the Cronbach’s alpha coefficient for the post-test results of the two AI concepts tests and the AI literacy, empowerment and ethical consideration surveys were 0.66, 0.67, 0.89, 0.93 and 0.76, respectively.

Focus group interviews were held to solicit in-depth views from the participants on the content of the courses and their perceptions of AI’s relevance to society. The participants were asked to write a reflective piece on their understanding of AI and related ethical issues before and after joining the courses. Besides employing the survey, the study evaluated participants’ ethical awareness by analysing their self-reflective writings. The text was analysed by a bilingual text-mining system using a keyword framework on AI and ethics (Kong et al., 2018; Kong, 2021; Kong et al., 2021a). To identify the level of participants’ ethical awareness, they were asked to write a self-reflective essay on ‘understanding of AI and ethics’ before and after Course 3, using the Moodle discussion forum. The participants were allowed to write in English or Chinese. Accordingly, a bilingual taxonomy of keywords with synonyms in both English and Chinese (see Table 2) was designed. To do so a course instructor and a research staff independently went through contents of the courses and the reflective essays of the participants to identify keywords, followed by discussions to arrive at the final version. We then used the bilingual text-mining system to count the number of keywords in the participants’ self-reflective writing.

4. Results and discussion

This section reports the results of the two AI concepts tests and the AI literacy, empowerment and ethical consideration surveys both before and after attending the courses. The key findings from the self-reflective writings are also included below. The discussions here complement those in two related publications: Kong and Zhang (2021) discuss thoroughly the establishment of the AI literacy framework outlined in Section 2, whereas Kong et al. (2021b) reports findings about Course 1. The current discussion, in contrast, offers a longitudinal investigation of students completing the entire programme.

4.1. Developing AI concepts and literacy

This section reports the development of the participants’ conceptual understanding and AI literacy. The results of the AI concepts tests and AI literacy survey show that the three courses successfully enhanced the participants’ conceptual understanding and literacy. Tables 3 and 4 show the means, standard deviations and paired *t*-test scores of the first and second concepts tests, respectively. The findings show that the increase in learning achievements in both concepts tests was statistically significant. This indicates that course participants from diverse backgrounds experienced significant progress in grasping AI-related concepts. This also implies that Course 1 and Course 2 provided participants with conceptual readiness for developing AI applications, which also built a framework for discussing ethics.

Table 3. Statistical results on the AI concepts Test 1 before and after Course 1

Concept	Before Course 1 (max. mark = 14)		After Course 1 (max. mark = 14)		Paired <i>t</i> -test
	Mean	<i>SD</i>	Mean	<i>SD</i>	
Machine learning	6.87	2.00	10.75	2.20	9.49***

Note. *N* = 36; **p* < .05; ***p* < .01; ****p* < .001.

Table 4. Statistical results on the AI concepts Test 2 before and after Course 2

Concept	Before Course 2 (max. mark = 14)		After Course 2 (max. mark = 14)		Paired <i>t</i> -test
	Mean	<i>SD</i>	Mean	<i>SD</i>	
Deep learning	6.72	2.50	9.19	2.75	4.68***

Note. *N* = 36; **p* < .05; ***p* < .01; ****p* < .001.

Participants also reported whether they had programming knowledge. Table 5 further compares how participants with and without programming knowledge performed in the concepts' tests. The results show that the two groups of participants did not exhibit statistically significant difference before Course 1, after Course 1 and before Course 2. However, after Course 2 participants without programming knowledge demonstrated even better performance which is statistically significant. This analysis supports that our courses, while not involving programming, are suitable for participants from diverse backgrounds to develop AI concepts.

Table 5. Comparing results of AI concepts tests by participants with and without programming knowledge

Concepts Test 1 (max. mark = 14)	Without programming knowledge (N = 14)		With programming knowledge (N = 22)		Paired <i>t</i> -test
	Mean	SD	Mean	SD	
Before Course 1	7.29	1.98	6.59	2.02	1.02
After Course 1	11.36	2.37	10.36	2.04	1.34
Concepts Test 2 (max. mark = 14)					
Before Course 2	7.14	2.03	6.45	2.77	0.80
After Course 2	10.57	2.47	8.32	2.61	2.58*

Note. N = 36; **p* < .05; ***p* < .01; ****p* < .001.

The participants' perceptions of their own AI literacy (see Table 6) increased significantly and stabilised after Course 1. This may be attributable to the ceiling effect, as a similar phenomenon was witnessed by Lee et al. (2021). The mean scores remained high after Course 1 (with the mean score above 4.10 out of 5).

Table 6. Statistical results on the AI literacy survey before and after the courses

	Before Course 1 Mean (SD)	After Course 1 Mean (SD)	After Course 2 Mean (SD)	After Course 3 Mean (SD)	<i>F</i> - value	<i>p</i> -value	Partial eta squared	Pairwise comparison
AI literacy (max. mark = 5)	2.74 (0.72)	4.10 (0.41)	4.07 (0.46)	4.19 (0.41)	46.91	< .001***	0.81	Before Course 1 < After Course 1; Before Course 1 < After Course 2; Before Course 1 < After Course 3;

Note. N = 36; **p* < .05; ***p* < .01; ****p* < .001.

Table 7. Selected quotes from participants' interviews and self-reflective writing after Course 3 on the usefulness of the project work in developing AI concepts

In terms of learning the concepts, my understanding of AI was not clear until conducting the project work. Now I understand that the study of AI focuses on computers' abilities to learn and identify objects. In our project ("Mask detection, home security and reminder systems"), the application identified whether the targets wore masks and if they were family members. This application development process helped me grasp the key elements of AI, which furthered my conceptual understanding (interview; S5).

The project work in Class 3 was indeed practical. We had opportunities to exchange ideas with others. The project work allowed us to present in class, interact with each other and discuss the topics and brainstorm how to implement the AI programme. These interactions and hands-on experiences contributed significantly to my understanding of AI concepts (interview; S14).

I used the "teachable machine" concept in my project. Through the project work, I deeply understood the underlying rationale of the algorithm. My conceptual understanding was strengthened so that in my daily life, I am now aware of the working principles of the algorithms behind some common related AI artefacts (interview; S31).

After Course 3, I have a better understanding of AI concepts, principles and applications. Although the trial phase was full of challenges, from trial and observation, we have learned about AI practices in society. This makes me more eager to apply the principles of AI in my work, and at the same time I am eager to improve the application of AI. (reflection translated from Chinese; S1).

Beyond the significant improvement demonstrated in the test and survey, participants' responses in focus group interviews and their reflective pieces (see Table 7) after Course 3 validated the usefulness of the project work for their understanding of AI concepts. For instance, some of the participants mentioned that they better understood AI concepts after completing their projects. They understood that AI involves computer learning and object

identification, an analytical process undertaken before making decisions. They acquired knowledge of the underlying principles of algorithms more thoroughly, which strengthened their conceptual understanding. Thematic analysis of interview transcripts and reflective writings showed the project works deepened participants' knowledge of machine learning and various platforms for developing AI applications (see Figure 3). This validated the feasibility of teaching concepts through project work, transforming their conceptual understanding from simple knowledge acquisition to novel applications of that knowledge (Roth, 1990; Schleicher, 2018). This in turn contributed to deeper conceptual understanding.

The development of participants' AI literacy was also evidenced by the results of their projects. Table 8 provides an overview of the 10 projects with project names and the total score and sub-score for ethical considerations each group received. In the project work, each group defined a real-life problem and created an application using the platforms demonstrated in class to solve that problem. In this process, they also evaluated their application by considering the possible ethical issues involved, gaining a better understanding of the permeation of AI in the real world (Kong & Zhang, 2021).

Table 8. Overview of group projects on “Developing Artificial Intelligence Applications”

Project no.	Project name	Number of participants	Total score (max.: 12)	Score for discussion of ethical considerations (max.: 3)
1	“Mind-RoadBot”: A mindfulness chatbot	4	12.00	3.00
2	“Perfect Letter”: A tool to help toddlers learn good English handwriting	2	10.50	2.36
3	Grade prediction model for evaluating students' learning outcomes	4	10.00	2.40
4	“Play with Sol-Fa Names”: Recognising hand signals for musical notes	3	10.00	2.00
5	Mask detection, home security and reminder systems	3	8.67	2.17
6	Predicting stock price to make a long-short portfolio	4	8.67	1.67
7	Garbage classification: Distinguishing different recyclable wastes from photos	5	8.25	2.17
8	“General Education Helper”: A Q&A bot for choosing general education courses	4	7.67	1.50
9	“Healthy Life Helper”: A chatbot to share health-related tips	3	7.50	1.84
10	A chatbot for song recommendations	4	6.50	1.88

Note. The total score is an average based on the marks from the instructors and participants.

4.2. Developing AI empowerment

Table 9 shows the results of the AI empowerment survey. As indicated, the mean scores remained high throughout the programme (above 4 marks a maximum score of 5) but increased significantly. This shows that participants felt empowered by learning the concepts and acquiring experience in developing AI applications.

Tracking the changes in participants' perceptions of their level of AI empowerment by course suggests that application development empowered the participants. The mean scores after Course 1 remained at a high level with no significant increase. This could be caused by the already high pre-course score (4.02 of 5), which signals that most participants in this group felt highly empowered even before starting the programme. Lee et al. (2021) reported a similar phenomenon in a study of AI literacy education and noted that this could be attributable to the ceiling effect. Although a significant decrease in the average score was observed before and after Course 2 (from 4.20 to 4.06), the mean score remained high. This may be due to the greater demands placed on participants in Course 2 as compared to Course 1. Student S4, for example, expressed in her reflective writing after Course 2 that “...I found that the knowledge behind is actually quite complicated...” and “...we have to handle different technical issues such as the overfitting problem.” This fluctuation in AI empowerment deviates from the results of the AI concepts tests which increased over both courses. One reason might be that the 36 students who ultimately completed the programme had high expectations for themselves and did not feel more empowered through the course, especially when their knowledge had yet to be applied. This also indicates the necessity of project work in application development. Tissenbaum et al. (2019) similarly argued that digital empowerment

involves instilling in learners the belief they can move beyond learning into meaningful action, suggesting project work as a means to achieve this. In this vein, our study engaged the participants in meaningful projects to empower them. This also explains the significant increase witnessed in the mean score before and after Course 3, which peaked at 4.22. The project work provided the participants with a valuable experience to help them creatively apply their knowledge to novel situations (Schleicher, 2018), further empowering the participants and actualising their digital creativity (Lee & Chen, 2015). These results are in line with the goal of cultivating AI-empowered, proactive citizens (Pemberton et al., 2019), who can leverage the benefits of AI and contribute to society more generally (JRC & OECD, 2021). In future offering of our programme, the level of difficulty of Course 2 is to be adjusted to suit the participants' needs better for enhancing AI empowerment. Considering the importance of project work, more reference to real-life applications can also be added to Course 2.

Table 9. Statistical results on the AI empowerment survey before and after the courses

	Before Course 1	After Course 1	After Course 2	After Course 3	<i>F</i> -value	<i>p</i> -value	Partial squared eta	Pairwise comparison
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)				
AI empowerment (max. mark = 5)	4.02 (0.49)	4.20 (0.49)	4.06 (0.50)	4.22 (0.40)	3.96	< .05*	0.27	After Course 1 > After Course 2; After Course 2 < After Course 3; Before Course 1 < After Course 3

Note. *N* = 36; **p* < .05; ***p* < .01; ****p* < .001.

4.3. Developing ethical awareness around AI

As the focus of the programme, the participants' development in ethical awareness around AI was demonstrated both through the surveys and their self-reflective writing. Their performance in the project work further validated the growth in their awareness. Table 10 shows the means, standard deviations and paired *t*-test scores of the ethical consideration survey before and after Course 3, demonstrating a statistically significant increase. This shows that the participants' perceived level of their own ethical awareness was enhanced.

Table 10. Statistical results on the ethical consideration survey before and after Course 3

Ethical consideration	Before Course 3 (max. mark = 5)		After Course 3 (max. mark = 5)		Paired <i>t</i> -test
	Mean	<i>SD</i>	Mean	<i>SD</i>	
	4.07	0.36	4.22	0.37	

Note. *N* = 36; **p* < .05; ***p* < .01; ****p* < .001.

The statistical results for counting the matched keywords related to AI ethics are shown in Table 11. The statistically significant increase in the mean scores demonstrates that the participants made significant progress in understanding real-world AI examples and principles.

Table 11. Statistical results for counting matched keywords related to AI ethics based on participants' self-reflections on AI ethics before and after Course 3

Keyword category	Before Course 3 reflection on AI ethics		Before Course 3 reflection on AI ethics		Paired <i>t</i> -test
	Mean	<i>SD</i>	Mean	<i>SD</i>	
	0.33	0.72	3.28	2.24	
Real-world examples	0.33	0.72	3.28	2.24	7.88***
Principles	0.22	0.49	0.89	1.39	2.58**
All keywords	0.56	0.88	4.17	2.57	8.10***

Note. *N* = 36; **p* < .05; ***p* < .01; ****p* < .001.

The increase in participants' ethical awareness was also reflected in their project work. The ethical considerations involved in each project are listed by project in Table 12. The participants considered the possible ethical implications emerging from the design, deployment and use of the application at the initial design stage.

Table 12. Ethical considerations in group projects on “Developing Artificial Intelligence Applications”

Project no.	Ethical considerations
1	<ul style="list-style-type: none"> • Detect and remove hate speech and discriminatory, embarrassing or biased answers. • Protect the privacy of users’ personal data; avoid asking about users’ private matters. • Avoid asking users questions containing marketing messages.
2	<ul style="list-style-type: none"> • Require consent to collect images of English letters from others. • Accuracy of the results affects toddlers’ engagement and teaching effectiveness.
3	<ul style="list-style-type: none"> • Some people may not possess the knowledge to check the validity of AI algorithms. • Be cautious of how to interpret the results generated by AI; regulate the parties accountable for making decisions with AI. • Require the consent of students and parents for data sharing and data exploration.
4	<ul style="list-style-type: none"> • Require consent to collect images of hand signals from others. • Accuracy of the system influences children’s learning motivation.
5	<ul style="list-style-type: none"> • Privacy issues may arise in labelling people other than family members as friends or strangers. • Ensure data safety. • Safeguard home users’ privacy without sacrificing too much freedom.
6	<ul style="list-style-type: none"> • Some people may not possess the knowledge to check the validity of AI algorithms. Information inequality arises. • There is uncertainty over who is held accountable for algorithm-based investment recommendations.
7	<ul style="list-style-type: none"> • Collect authorised photos to train the AI model. • Implement sufficient data protection policies; avoid sharing AI inference results for commercial purposes.
8	<ul style="list-style-type: none"> • Protect the data privacy of users’ conversation. • Be aware of the potential discrimination portrayed in the chatbot’s responses.
9	<ul style="list-style-type: none"> • Protect the privacy of user’s body measurement data; keep user informed of how personal data are processed.
10	<ul style="list-style-type: none"> • Avoid potential bias in the music database, e.g., whether to include commercial music. • Protect the data privacy of users’ conversation.

Table 13. Selected quotes from participants’ interviews and self-reflections after Course 3 on the usefulness of the project work in developing their ethical awareness

Course 3 enhanced my critical thinking skills and awareness of AI ethics. While collecting data for our project, I also reflected on the ethical debate concerning AI and gained a deeper understanding of whether the advantages outweighed the disadvantages or vice versa. This process increased my critical thinking and ethical awareness (interview; S7).

My ethical awareness was enhanced and developing such an awareness interests me a lot. As a novice language teacher, developing the general analytical skills and thinking through the ethics of AI are even more important than developing an AI model. It is important to integrate technologies in teaching, but what matters even more is ethical awareness. This enhanced my analytical skills in deciding which data to use, considering copyright and other ethical issues, which all benefit our individual development. Today, being aware of how to use AI - knowing how to cope with societal change - really matters for both teachers and students. I benefited a lot from the project work (interview; S11).

The project work in Course 3 prompted me to reflect on ethics. Before attending the courses, I thought through application development with no concern for ethics. The two cases studies in the beginning of Course 3 served as effective examples of ethical issues in application development. This course reminded me to consider the possible ethical implications involved (interview; S18).

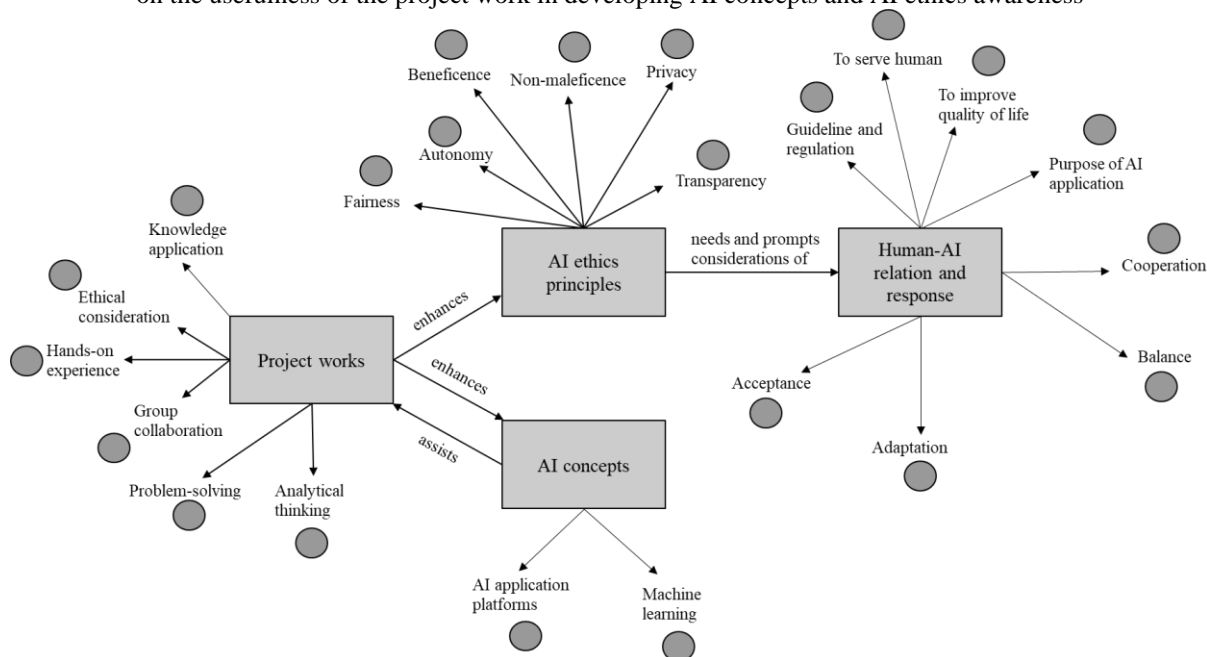
Many machine learning models generate their results by operating on high-dimensional correlations beyond the interpretive capabilities of human-scale reasoning. In these cases, the rationale of algorithmically produced outcomes that directly affect decision subjects remains opaque to those subjects. In some applications, the processed data could cause discrimination, bias, inequity or unfairness. The opaqueness of the model may be deeply problematic. Therefore, people should pay attention to those issues before applying AI techniques (reflection; S14).

Figure 3 reports the thematic analysis of participants’ reflective pieces and focus group interviews. Four themes were identified, namely “project works,” “AI concepts,” “AI ethics principles” and “human-AI relation and response.” In the thematic map, squares represent the themes, with the codes, represented by circles, emanating

from the corresponding themes via arrows. Table 13 shows sample quotations and Appendix 1 shows the operational definition of each code.

Participants reported that the AI application development projects enabled them to apply their knowledge, have hands-on experience, consider ethical issues and train their analytical thinking, which collectively improved their understanding and awareness of AI ethical principles. Not only did participants consider principles explored in the course (namely autonomy, beneficence/non-maleficence and fairness), they also mentioned elements of AI ethics beyond the curriculum, such as transparency and privacy. Furthermore, participants were able to provide suggestions and considerations on human-AI relations and on humans' responses to the societal changes caused by AI. The finding fits with our goal of a holistic cultivation of AI literacy, which includes the ability to evaluate and reflect on AI in real-world scenarios. It also validates the feasibility and success of the novel approach to teaching AI ethics by integrating ethical considerations into project work. This approach differed from the delivery of abstract principles of AI ethics (Borenstein & Howard, 2021) as it effectively guided participants to reflect on the complex ethical concerns emerging from the design, deployment and use of AI technologies.

Figure 3. Thematic analysis of the text from participants' interviews and self-reflective writings after Course 3 on the usefulness of the project work in developing AI concepts and AI ethics awareness



5. Conclusions and implications

This study presented an evaluation of an AI literacy programme which developed university students' concepts, literacy, empowerment and ethical awareness. We conducted a 30-hour programme and piloted it with 36 university students in Hong Kong with diverse backgrounds. Our surveys, tests and self-reflective writing assignments demonstrated that the course participants felt empowered and made significant gains in their understanding of major AI concepts, literacy and ethical awareness.

One limitation of our programme is the decline in participant number from 120 in Course 1 to 82 in Course 2 and to 36 in Course 3, which may be attributed to it being non-credit-bearing and participants' other credit-bearing activities (Oakley et al., 2011). Despite this decline, participants' evaluation of each course showed their satisfaction, even when including outgoing participants, suggesting the dropout was unrelated to course quality. Another limitation is the online teaching mode under COVID-19, which potentially affected participants' learning and participation (Salas-Pilco et al., 2022).

Despite the limitations, the results of our course have several important implications. First, the study refocuses AI literacy programmes on conceptual building instead of first emphasising mathematical formulae and programming codes (Kong et al., 2021b). Teaching concepts in this way can lower the barrier and ensure equal access to AI literacy for people from all walks of life (Long & Magerko, 2020), which is a great leap in promoting AI literacy among educated citizens of diverse backgrounds.

Second, the study highlights the importance of project work in an AI literacy programme. The survey results demonstrated that Course 3 played a significant role in empowering participants. The project work unleashed their digital creativity (Lee & Chen, 2015) and deepened their conceptual learning by enabling them to creatively apply their knowledge to new contexts (see Section 4.1).

Finally, the study initiates and validates the method of teaching AI ethics through project work in application development. This approach differed from the delivery of abstract principles of AI ethics in an after-the-fact manner, and instead emphasised their importance at every stage of learning about AI (Borenstein & Howard, 2021). It effectively guided the participants to reflect on the complex ethical concerns emerging from the design, deployment and use of AI technologies.

The significance of the study lies in its validation of a pathway to develop AI literacy among educated citizens from diverse academic backgrounds. It not only contributes to the demystification of AI among the public by fostering conceptual understanding, but also cultivates AI-empowered, proactive and ethically informed citizens who can leverage the benefits of AI to contribute to society more generally.

Acknowledgement

The work described in this paper was substantially supported by a grant from the Research Grants Council, University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. EdUHK CB302), and received funding support from the Li Ka Shing Foundation.

References

- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for learning, teaching, and assessing: A Revision of Bloom's taxonomy of educational objectives*. Longman.
- Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and digital technologies. *International Journal of Information Management*, 62, 102433. <https://doi.org/10.1016/j.ijinfomgt.2021.102433>
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122-147. <https://doi.org/10.1037/0003-066X.37.2.122>
- Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, 1(1), 61-65. <https://doi.org/10.1007/s43681-020-00002-7>
- European Commission, Joint Research Centre (JRC) & Organisation for Economic Co-operation and Development (OECD). (2021). *AI watch, national strategies on artificial intelligence: A European perspective*. Publications Office of the European Union. <https://doi.org/10.2760/069178>
- Fosso Wamba, S., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D. A. (2021). Are we preparing for a good AI society? A Bibliometric review and research agenda. *Technological Forecasting and Social Change*, 164, 120482. <https://doi.org/10.1016/j.techfore.2020.120482>
- Frymier, A. B., Shulman, G. M., & Houser, M. (1996). The Development of a learner empowerment measure. *Communication Education*, 45(3), 181-199. <https://doi.org/10.1080/03634529609379048>
- Floridi, L., & Cowls, J. (2019). A Unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), <https://doi.org/10.1162/99608f92.8cd550d1>
- Garrett, N., Beard, N., & Fiesler, C. (2020). More than “If time allows”: The Role of ethics in AI education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 272-278). <https://doi.org/10.1145/3375627.3375868>
- Green, N. (2021). An AI ethics course highlighting explicit ethical agents. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 519-524). <https://doi.org/10.1145/3461702.3462552>
- High Level Expert Group on Artificial Intelligence (HLEG). (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. <https://doi.org/10.2759/177365>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The Global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kong, S. C., Li, P., & Song, Y. (2018). Evaluating a bilingual text-mining system with a taxonomy of key words and hierarchical visualization for understanding learner-generated text. *Journal of Educational Computing Research*, 56(3), 369-395. <https://doi.org/10.1177/0735633117707991>

- Kong, S. C. (2021). Delivery and evaluation of an e-learning framework through computer-aided analysis of learners' reflection text in a teacher development course. *Research and Practice in Technology Enhanced Learning*, 16, Article 28. <https://doi.org/10.1186/s41039-021-00172-w>
- Kong, S. C., Kwok, W.Y., & Poon, C.W. (2021a). Evaluating a learning trail for academic integrity development in higher education using bilingual text mining. *Technology, Pedagogy and Education*, 30(2), 305-322. <https://doi.org/10.1080/1475939X.2021.1899041>
- Kong, S. C., Cheung, W. M. Y., & Zhang, G. (2021b). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence*, 2, 100026. <https://doi.org/10.1016/j.caeai.2021.100026>
- Kong, S. C. & Zhang, G. (2021). A Conceptual framework for designing artificial intelligence literacy programmes for educated citizens. In S. C. Kong, Q. Wang, R. Huang, Y. Li, & T. C. Hsu (Eds.), *Conference Proceedings (English Track) of the 25th Global Chinese Conference on Computers in Education, GCCCE 2021* (pp. 11-15). The Education University of Hong Kong.
- Kuipers, B. (2020). Perspectives on ethics of AI: Computer science. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 419-441). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190067397.013.27>
- Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021). Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 191-197). <https://doi.org/10.1145/3408877.3432513>
- Lee, M. R., & Chen, T. T. (2015). Digital creativity: Research themes and framework. *Computers in Human Behavior*, 42, 12-19. <https://doi.org/10.1016/j.chb.2014.04.001>
- Long, D. & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-16), Honolulu, United States. <https://doi.org/10.1145/3313831.3376727>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35. <https://doi.org/10.1145/3457607>
- National Commission for the Protection of Human Subjects (NCPHS). (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.
- Oakley, G., Lock, G., Budgen, F., & Hamlett, B. (2011). Pre-service teachers' attendance at lectures and tutorials: Why don't they turn up? *Australian Journal of Teacher Education*, 36(5), 31-47. <http://dx.doi.org/10.14221/ajte.2011v36n5.3>
- Organisation for Economic Co-operation and Development (OECD). (2018a). *Bridging the digital gender divide: Include, upskill, innovate*. OECD Publishing. <http://www.oecd.org/digital/bridging-the-digital-gender-divide.pdf>
- Organisation for Economic Co-operation and Development (OECD). (2018b). *Future of education and skills 2030: Conceptual learning framework*. OECD Publishing. <https://www.oecd.org/education/2030/Education-and-AI-preparing-for-the-future-AI-Attitudes-and-Values.pdf>
- Organisation for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. OECD Publishing. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Paulus, P. B., & Brown, V. R. (2003). Enhancing ideational creativity in groups: Lessons from research on brainstorming. In P. B. Paulus & B. A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 110-136). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195147308.003.0006>
- Pemberton, D., Lai, Z., Li, L., Shen, S., Wang, J., & Hammer, J. (2019). AI or Nay-I? Making moral complexity more accessible. In *Extended abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion* (pp. 281-286). <https://doi.org/10.1145/3341215.3358248>
- Pouly, M., Koller, T., & Arnold, R. (2019). A Game-centric approach to teaching artificial intelligence. In *Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019)* (pp. 398-404). <https://doi.org/10.5220/0007745203980404>
- Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*, 4(2), 99-101. <https://doi.org/10.1038/s42256-022-00449-9>
- Roth, K. (1990). Developing meaningful conceptual understanding in science. In B. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 139-175). Lawrence Erlbaum Associates, Inc.
- Salas-Pilco, S. Z., Yang, Y., & Zhang, Z. (2022). Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A Systematic review. *British Journal of Educational Technology*, 53(3), 593-619. <https://doi.org/10.1111/bjet.13190>

- Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, 19(4), 1-26. <https://doi.org/10.1145/3341164>
- Schleicher, A. (2018). *World class: How to build a 21st-century school system: Strong performers and successful reformers in education*. OECD Publishing. <https://doi.org/10.1787/9789264300002-en>
- Stadelmann, T., Keuzenkamp, J., Grabner, H., & Würsch, C. (2021). The AI-Atlas: Didactics for teaching AI and machine learning on-site, online, and hybrid. *Education Sciences*, 11(7), 318. <https://doi.org/10.3390/educsci11070318>
- Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I. & Pears, A. (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE Access*, 9, 110558-110572. <https://doi.org/10.1109/ACCESS.2021.3097962>
- Thomas, K. W., & Velthouse, B. A. (1990). Cognitive elements of empowerment: an “interpretive” model of intrinsic task motivation. *Academy of Management Review*, 15(4), 666-681. <https://doi.org/10.2307/258687>
- Tissenbaum, M., Sheldon, J., & Abelson, H. (2019). From computational thinking to computational action. *Communications of the ACM*, 62(3), 34-36. <https://doi.org/10.1145/3265747>
- Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9795–9799. <https://doi.org/10.1609/aaai.v33i01.33019795>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The Role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>
- Wong, G. K. W., Ma, X., Dillenbourg, P., & Huan, J. (2020). Broadening artificial intelligence education in K-12: Where to start? *ACM Inroads*, 11(1), 20-29. <https://doi.org/10.1145/3381884>
- World Intellectual Property Organization (WIPO). (2019). *WIPO technology trends 2019 – Artificial Intelligence*. World Intellectual Property Organization. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf
- Yi, Y. (2021). Establishing the concept of AI literacy: Focusing on competence and purpose. *Jahr - European Journal of Bioethics*, 12(2), 353-368. <https://doi.org/10.21860/j.12.2.8>

Appendix 1. Codes and definitions for thematic analysis

Theme	Code	Operational definition
AI Ethics Principles	Fairness	AI applications do not have prejudice, bias or stereotypes to any individuals, and their benefits and harms are equally distributed to everyone
	Beneficence	AI applications actively promote humanity's safety and well-being
	Non-maleficence	AI applications do not harm humans
	Transparency	The usage, merits and drawbacks of AI applications are clearly stated
	Autonomy	AI applications are not used to manipulate people. Humans are the ones to make decisions and be accountable
	Privacy	AI applications protect the security of people's data and do not infringe on people's privacy
Human-AI Relation and Response	Adaptation	People adapt to a world permeated with AI
	Acceptance	People accept the changes brought about by AI
	Balance	People balance the risks and benefits of AI before adopting it
	Guideline and regulation	People regulate the usage of AI and establish guidelines to manage it
	To improve quality of life	AI improves people's quality of life
	To serve human	AI serves humans as a tool
	Cooperation	AI cooperates with people and complements people's shortcomings
	Purpose of AI application	People think about the purpose of using AI before adopting it
Projects	Knowledge application	Participants apply their knowledge acquired previously
	Hands-on experience	Participants practice their knowledge hands-on
	Group collaboration	Participants collaborate and interact with groupmates
	Ethical consideration	Participants consider the ethics of their AI applications
	Problem-solving	Participants solve real-world problems with their AI applications
	Analytical thinking	Participants think analytically and critically when evaluating their AI applications
AI Concepts	AI application platforms	Participants use platforms (such as Google Teachable Machine and Microsoft Azure QnA Maker) to develop their AI applications
	Machine learning	Participants apply their knowledge of machine learning learnt in the first 2 courses

A Systematic Review of Technology-Enhanced Self-Regulated Language Learning

Yin Yang¹, Yun Wen² and Yanjie Song^{1*}

¹Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong

// ²National Institute of Education, Nanyang Technological University, Singapore // yyin@s.eduhk.hk //

yun.wen@nie.edu.sg // ysong@eduhk.hk

*Corresponding author

(Submitted September 23, 2021; Revised April 16, 2022; Accepted April 25, 2022)

ABSTRACT: The role of self-regulated learning in language learning has been widely acknowledged, and there is a growing number of studies on technology-enhanced self-regulated language learning (SRLL). This systematic review aims to provide a holistic picture of existing studies in this area by identifying the characteristics of published studies, the research methods used to evaluate SRLL effectiveness and the role of technology in SRLL. The review covered 34 empirical studies focusing on SRLL that were published from 2011 to 2020. The results showed varied characteristics of technology-enhanced SRLL studies, dominance of the use of quantitative methods, greater focus on examining students' SRLL outcomes instead of their processes, and the role of technology in supporting the performance phase of students' SRLL instead of the entire SRLL process. These findings have implications for using technologies to facilitate and examine the holistic process of students' SRLL.

Keywords: Systematic literature review, Technology, Self-regulated language learning (SRLL)

1. Introduction

Online learning systems, especially mobile applications, are widely used in many educational contexts, including language teaching and learning. The boundaries between formal and informal language learning, classroom-based learning and out-of-classroom learning activities have become blurred and interconnected with the rapid development of wireless communication networks and mobile devices (Sharples et al., 2016). As this new environment provides unprecedented opportunities for language learning, learners should develop self-regulated learning (SRL) skills to succeed. They must set goals and schedule efficiently while participating in online learning activities (Yeh et al., 2019; Zhou & Wei, 2018).

SRL refers to an active, constructive process through which learners set learning goals and then attempt to monitor, regulate, and control their cognitive and metacognitive process and learning behaviours (Pintrich, 2000). It is also an essential component of lifelong learning to cope with the challenges of the twenty-first century (Lehmann et al., 2014; Zheng et al., 2018). Many studies have shown that SRL is positively related to students' learning outcomes (Chen et al., 2014; Lai et al., 2018). Zimmerman (2002) posited that self-regulatory processes are teachable. To improve learning outcomes, students must engage in effective SRL processes in planning and setting goals, monitoring their learning process and evaluating their whole learning performance (e.g., Azevedo et al., 2018; Lai et al., 2018). Interventions are necessary to support students in developing SRL (Yang et al., 2018; Yeh et al., 2019). In recent years, the number of studies on SRL in online learning environments has soared. In these works, researchers have focused on trends in measurement and intervention tools for SRL (Araka et al., 2020), the correlation between SRL strategies and academic achievement in online higher education (Broadbent & Poon, 2015), approaches to supporting SRL in online learning (Wong et al., 2018), the relationship between SRL and mobile learning (Palalas & Wark, 2020) and the relationship between SRL and learning analytics in online learning (Viberg et al., 2020). However, reviews of technology-assisted self-regulated language learning (SRLL) are scarce.

Preliminary studies in the field of language learning have investigated SRLL mediated by technologies, such as in reading (e.g., Chen et al., 2014; Zheng et al., 2018; Serrano et al., 2018), writing (e.g., Ducasse & Hill, 2019) and vocabulary learning (e.g., Chen & Hsu, 2020). By contrast, papers on the learning effectiveness of SRLL have had various foci, such as language learning outcomes (Chen et al., 2019), SRL strategies (Chen & Lee, 2018) and SRL skills (Yeh et al., 2019). A number of studies have shown that technology-enhanced learning environments can provide technological affordances for improving language learning outcomes and fostering SRL skills (Hromalik & Koszalka, 2018; Shyr & Chen, 2018; Woottipong, 2022). According to other studies, technology is not positively related to language learning outcomes (e.g., Chen & Lee, 2018) or SRL skills (e.g.,

Seifert & Har-Paz, 2020). This inconsistency in findings may be caused by the design of technology-assisted learning environments. A well-designed technology-enhanced learning environment can help learners regulate their learning, determine where and when to learn, cultivate their SRL behaviours and sustain their interest in SRL (Shih et al., 2010).

Therefore, in addition to exploring the characteristics of empirical SRL studies in terms of the publication years and learner types, this review study explored how SRL effectiveness was investigated and the role of technology in these SRL studies. To understand the trends of SRL in language learning and the potential of using technology to cultivate language learners' SRL skills and improve their learning performance, this systematic review examined technology-enhanced SRL studies published in the past 10 years with the following questions:

RQ1: What were the characteristics of SRL studies in terms of their publication years and learner types?

RQ2: What research methods were adopted to examine SRL effectiveness in the reviewed studies?

RQ3: What role did technology play in supporting SRL in the reviewed studies?

2. Method

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was applied to guide this systematic review (Moher et al., 2015) to ensure the rigour and quality of the review process. The search strategy, selection criteria and data coding and analysis in this review are presented below.

2.1. Search strategy

First, the major relevant terms used in the literature, including synonyms and alternative spellings, were identified. The following search string was then used to search for relevant articles: ("self-regulated" OR "self-regulatory" OR "self-regulation") AND ("language learning" OR "reading" OR "writing" OR "speaking" OR "grammar" OR "vocabulary") AND ("technology" OR "computer" OR "mobile" OR "tablet" OR "phone"). The data for this study were selected from the following academic journal databases: Educational Resources Information Centre (ERIC), Web of Science (WOS), Wiley and ProQuest. These databases are widely used in educational studies (Bano et al., 2018; Lee, 2019; Lin & Lin, 2019). The search only involved peer-reviewed articles that could be retrieved online to ensure a high quality of the selected articles (Hung et al., 2018). Endnote was used to track each identified citation and to manage and document the imported databases throughout the search process.

2.2. Selection criteria

Inclusion and exclusion criteria were applied to eliminate irrelevant studies. As illustrated in Table 1, the study had to be (1) published in English, (2) dated from 2011 to 2020 inclusively, (3) an empirical or case study, and (4) in a technology-enhanced language learning environment. Only articles from peer-reviewed journals were selected. Other types of publications, such as theses, book reviews and conference papers, were not included. This criterion is widely used in other literature reviews to maintain a high quality of selected papers (e.g., Lin et al., 2019; Shadiev & Yang, 2020; Zainuddin et al., 2020; Zou et al., 2019). In addition, special needs education research, including studies involving participants with dyslexia, was eliminated for the following reasons. First, such studies are commonly excluded from literature reviews related to technology-assisted language learning environments (e.g., Bano et al., 2018; Zou et al., 2019). Second, special needs education should be approached carefully and that technologies for individuals with autism spectrum disorders or cognitive disabilities are complex and deserve further investigation. During our literature search, we observed an increase in the amount of attention paid to SRL with technology in the field of special education (e.g., Ben-Yehudah & Brann, 2019; Hughes et al., 2019).

The search of online databases resulted in 466 articles. A total of 345 articles remained after the removal of duplicates, and their titles and abstracts were scanned. Two researchers were involved in the selection process to avoid selection bias. All articles were examined by both researchers; researcher A found 136 articles to be relevant, and researcher B considered 122 articles relevant. The titles of these articles were documented in Excel by two authors and compared one by one. In total, 136 articles were selected for further analysis.

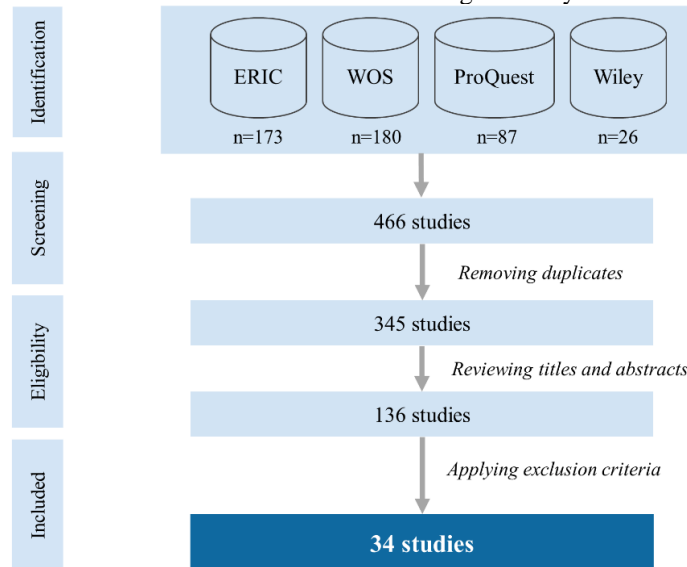
Table 1. Exclusion and inclusion criteria

Inclusion criteria	Exclusion criteria
<ul style="list-style-type: none"> Published in English Published from 2011 to 2020 Empirical studies and case studies In technology-enhanced language learning environments 	<ul style="list-style-type: none"> Published in other languages Not in technology-enhanced learning environments A thesis/editorial/book review/conference paper Inadequate information on research design and data analysis Literature review and conceptual studies in nature Special needs education research

The three researchers applied inclusion and exclusion criteria (see Table 1). The first author applied these criteria to all papers for study selection. The second author randomly checked the results by examining twenty papers. All questions related to article selection were resolved by the three authors together in a discussion. This process was guided by Bano et al. (2018) and Shadiev and Yang (2020). Finally, 34 papers were deemed eligible for the review. Among the cases included by Llorens et al. (2016) and Serrano et al. (2018) in their multi-case studies, only those related to this review were selected.

To sum up, the study selection process, which was based on PRISMA (Moher et al., 2015), is illustrated in Figure 1. A total of 34 articles (see Appendix A) were considered eligible for the review.

Figure 1. PRISMA flowchart of article screening in this systematic review



2.3. Data coding and analysis

All the selected papers were coded and analysed using content analysis. The first research question concerned the characteristics of the SRL studies' publication years and learner types. As for the publication years, the distribution of the selected papers in 2011–2020 was analysed. About the learner types, due to the weakness in metacognition of young learners (van Loon & Roebbers, 2017), SRL cultivation might be sensitive to age. Researchers have different opinions on whether children younger than six years can use metacognitive strategies (Dignath & Büttner, 2008). There are also studies suggesting that children aged 7 to 8 years self-evaluate less compared with those aged 11 to 12 years (Paris & Newman, 1990). Paris and Winograd (1999) state that children's metacognition develops during schooling from the age of 5 to 16 years (Paris & Winograd, 1999). On this basis, in addition to categorising learners in terms of educational levels, we distinguished lower and higher primary school students. Hence, learners were classified into six sub-categories: (1) 6 years old and below, (2) 7–9 years old, (3) 10–15 years old, (4) 16–18 years old, (5) undergraduate and/or postgraduate, and (6) workplace adult learners.

The research methods used to evaluate learning effectiveness were coded and analysed to address the second research question. These approaches were categorised as quantitative, qualitative, and mixed methods. The durations of the studies were classified into the following categories, which were adapted from Hwang and Fu (2019): one session, short term (< 10 weeks), intermediate term (11 weeks to 4 months) and long term (> 4

months). The evaluation of SRL was divided into four categories: (1) assessing student language learning outcomes, (2) assessing students' SRL (e.g., self-efficacy, attitudes, SRL strategies used, SRL skills behaviours) (Ardasheva et al., 2017; Panadero et al., 2016), (3) assessing both language learning outcomes and SRL, and (4) exploring the technology-enhanced SRL profiles of students.

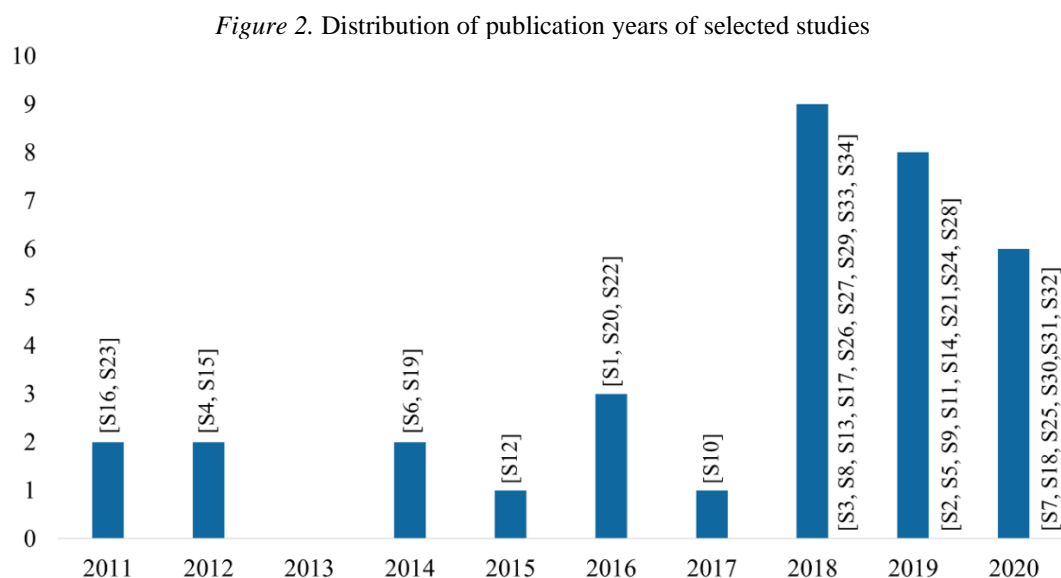
As for the third research question, first, technology was categorised as self-developed or third-party technology. The former referred to technology explicitly designed by researchers to investigate its use in teaching and learning, whereas the latter meant commercial software or technology that was developed by a third party. The types of technology were coded as mobile devices (e.g., mobile phones and iPads), desktop personal computers (PCs), and multiple devices. Multiple devices referred to the presence of more than one type of device in the study. Second, the learning settings, which referred to the contexts in which these technologies were employed, were divided into the following categories, which were based on Bano et al. (2018): formal settings, informal settings, multiple settings and not specified. Formal settings referred to traditional learning environments, such as institutionalised settings (e.g., public schools and universities); informal settings included learning spaces apart from formal learning settings, such as homes, subways, gardens and supermarkets; multiple settings referred to combinations of formal and informal learning experiences; not specified meant that no specific learning context was indicated in the study. Lastly, the role of technology in supporting SRL processes was coded in terms of Zimmerman's (2002) SRL model, which is widely acknowledged in the field (Dignath et al., 2008; Panadero, 2017). According to Zimmerman (2002), SRL processes consist of the following phases: forethought, performance and self-reflection. The forethought phase involves task analysis (e.g., goal setting and strategic planning). In the performance phase, students monitor their processes. Finally, the self-reflection phase includes self-judgment and self-reactions to learning performance and outcomes. These phases were used to analyse and address the third question.

In piloting the coding scheme, two researchers coded eight articles together and discussed the coding results until a consensus was reached. After that, the same two researchers independently coded the 26 remaining articles. Cohen's kappa, which was calculated to measure the inter-rater reliability about the role of technology in terms of Zimmerman's (2002) SRL model, was 0.91, which was considered perfect (Stemler, 2004). In finalising the coding results of the 34 selected articles, the three researchers discussed any discrepancy by conducting face-to-face discussions and by rechecking points of disagreement until a consensus was reached.

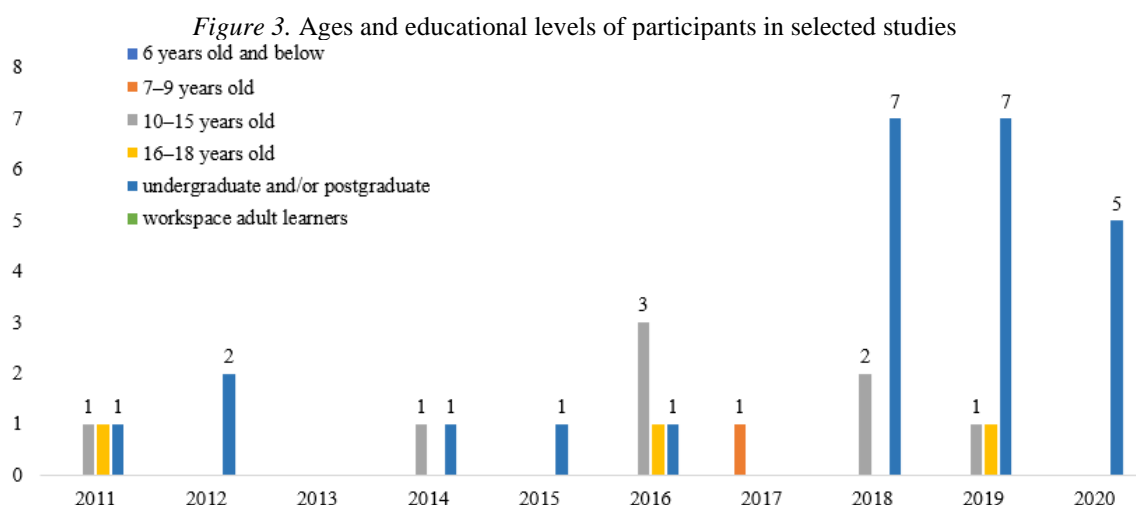
3. Results

3.1. Study characteristics in terms of the publication years and learner types

Figure 2 shows the distribution of the publication years of the 34 selected articles over the past 10 years (2011–2020). The number of empirical research papers dramatically increased from 2017 to 2018 but declined between 2018 and 2020.



As for the learner types, Figure 3 shows that approximately 73.5% of the studies were conducted in higher education. Eight studies involved participants between the ages of 10 and 15 years, followed by three studies conducted among participants between the ages of 16 and 18 years. Only one study focused on participants aged 7 to 9 years. No study involved participants aged below 6 years or workplace adults.



3.2. Research methods adopted to examine SRL effectiveness

The SRL learning effectiveness was examined in the selected studies via different research methods. Eighteen studies (53%) primarily adopted quantitative research methods and mainly aimed to investigate the effectiveness of developed mobile applications or learning systems for SRL. A total of 12 studies (35%) employed mixed research methods and four studies (12%) adopted qualitative research methods to explore students' SRL experience. Twelve papers (35%) were intermediate-term studies, and eight were short-term studies (23%). Six studies (18%) were conducted in one session each, whereas five (15%) were long-term (longer than four months) studies.

Among the 34 studies, 10 studies adopted non-experimental research designs and investigated students' self-regulated technological profiles. In these studies, students were free to choose and adopt various tools to regulate their language learning. Among these 10 studies, seven described how students used technology to regulate their language learning experience in online environments using closed-ended survey questionnaire (Tao et al., 2020), interviews (Lai & Gu, 2011; Lei, 2018; Wang & Chen, 2020), participant-made videos (Chien, 2019), open-ended survey questionnaires (Lai & Gu, 2011; Su et al., 2019) and reflective journals (Hromalik & Koszalka, 2018). In addition, six studies mainly employed questionnaires and correlational analysis to understand the underlying relationship between SRL factors.

The 24 remaining studies were conducted using experimental research designs. Eighteen studies (52.9%) investigated both language and SRL outcomes, whereas six studies (17.6%) only focused on self-regulation. Detailed information is presented in Table 2. Among the 24 studies, 18 (52.9%) investigated students' language learning outcomes. Quizzes were primarily adopted to assess students' improvement in language learning. Only one study used students' e-portfolios, where students recorded their oral production to assess the progress of their oral performance (Torres et al., 2020).

As for evaluating students' SRL, 20 studies (83.3%) used surveys, including questionnaires, self-reports and interviews. Six studies employed log data to analyse students' SRL-related behaviours while interacting with the studied technologies. For example, Chen et al. (2014) used data recorded on a digital reading annotation system (DRAS) of the achievement index of learning time, effort level, reading rate, concentrated learning and degree of understanding of learned courseware to assess students' SRL skills. Lee and Chan (2018) traced students' behaviours on the My-Pet-Shop system to explore SRL behavioural patterns. Kondo et al. (2012) analysed time spent on Nintendo DS mobile devices. Llorens et al. (2016) mainly analysed students' behaviours recorded online to indicate self-regulation strategies and decision-making; these recorded behaviours were the number of times students decided to revisit text or questions and the decisions made by the students at specific times. Roussel (2011) recorded students' physical movements of the mouse during a listening task to indicate their ability to regulate their listening in language learning. Serrano et al. (2018) measured students' monitoring

accuracy by calculating the number of right or wrong answers in non-search decisions to help them regulate their use of text information in reading. Moreover, four studies involved teachers' observation (Ferreira et al., 2017; Ghuftron & Nurdianingsih, 2019; Karami et al., 2019; Torres et al., 2020). Ferreira et al. (2017) considered teachers' perspectives by asking them to rate their students' SRL using a questionnaire. Ghuftron and Nurdianingsih (2019) used an observation protocol to document in-class teaching and learning for analysis. In the study of Karami et al. (2019), teachers' field notes were used to triangulate students' surveys in order to understand their SRL in English writing. Similarly, teachers' journals which included observations of students' performance were used by Torres et al. (2020) to explain students' strategy use in developing their English speaking skills.

Table 2. Summary of 24 studies assessing student language-related and/or SRL

Paper ID	Language learning outcomes								SRL			
	Focus						Data sources		SRL ability/strategy	Data source		
	V	W	M	R	S	L	Qz	Others		Survey	Log data	Observation
S1					x				x	x		
S3				x					x	x		
S5	x						x		x	x		
S6				x			x		x		x	
S7			x				x		x	x		
S8	x						x		x	x	x	
S10			x				x		x	x		x
S11		x							x	x		x
S12	x						x		x	x		
S14		x					x		x	x		x
S15			x				x		x		x	
S18		x					x		x	x		
S19	x						x		x	x		
S20			x				x		x		x	
S21		x							x	x		
S22						x			x	x		
S23						x	x		x		x	
S24			x						x			
S25			x				x		x	x		
S26				x			x		x		x	
S27	x						x		x	x		
S31					x			x	x	x		x
S33				x			x		x	x		
S34				x			x		x	x		

Note. V = vocabulary; W = writing; M = mixed language learning outcomes; R = reading; S = speaking; L = listening; Qz = quizzes.

3.3. Role of technology in supporting SRL

Table 3 presents the technologies, devices, tool descriptions and learning settings in the 34 selected papers. A total of 15 studies (44.1%) used self-developed applications or systems. The rest investigated the use of technologies developed by third parties. Regarding the adopted devices, 16 studies (47%) employed desktop PCs, 11 studies (32%) used mobile devices and seven studies (21%) adopted multiple devices. In addition, 10 studies (29%) employed free-to-use technology, but seven of them did not specify Web 2.0 technologies and applications (Al Fadda, 2019; Çelik et al., 2012; Chien, 2019; Hromalik & Koszalka, 2018; Lai & Gu, 2011; Su et al., 2018; Tao et al., 2020).

The majority of studies (55.9%) were conducted across multiple learning contexts (e.g., classrooms and homes). The rest occurred in formal settings (35.3%) or informal settings (Çelik et al., 2012; Lai & Gu, 2011; Zhai et al., 2018). More studies were conducted in informal settings in the first five years (2011–2015) than in the next five years (2016–2020). Overall, the distribution of learning contexts indicated that SRL research was conducted more frequently in multiple settings.

Table 3. List of technologies, devices, tools descriptions and learning settings in the selected studies

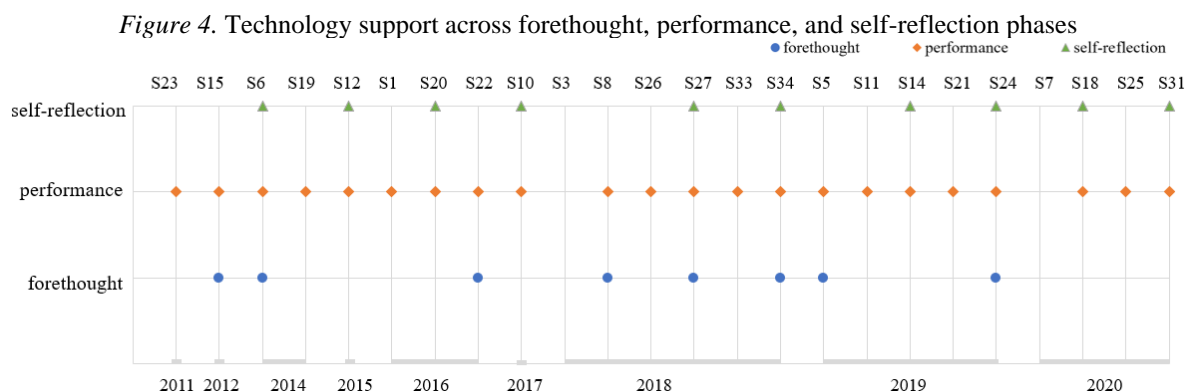
Paper ID	S	T	Devices	Tools description	Learning settings
S1	x		O	automatic speech recognition (ASR)	multiple
S2		x	M	multiple tools	multiple
S3		x	O	multiple tools: dictionaries, WhatsApp, camera, internet search engines, notes, and recorders	multiple
S4		x	M	multiple tools	informal
S5	x		O	EVLAPP-SRLM: an English vocabulary learning app with a self-regulated learning mechanism	formal
S6	x		D	a digital reading annotation system (DRAS): an SRL mechanism combined with useful annotation functionalities that can annotate digital texts in the HTML format	formal
S7	x		O	mobile virtual reality environment (VRE)	multiple
S8	x		D	My-Pet-Shop: an educational game to enhance young children's learning of English vocabulary	formal
S9		x	M	multiple tools	multiple
S10	x		D	an adapted Moodle platform	formal
S11		x	D	multiple tools: Grammarly, Google docs and Microsoft Word.	multiple
S12	x		O	a calibration scheme: using a preview or review process for individual learners	multiple
S13		x	M	multiple tools	multiple
S14		x	D	e-portfolio: Edmodo	multiple
S15		x	O	the Nintendo DS Lite: DS More Training for the TOEIC Listening and Reading Tests	multiple
S16		x	M	multiple tools	informal
S17		x	O	WeChat: free software provided by China mobile	multiple
S18	x		O	the ARCAUW application: using the software Unity for Mobile AR.	formal
S19		x	D	Google docs-Web-based collaboration tool	multiple
S20	x		D	Read&Answer: record students' search behaviour while reading	formal
S21		x	D	computer-mediated discussions	multiple
S22		x	D	podcast	formal
S23		x	D	recorders	formal
S24	x		D	the prompts added in the guidelines of the assignments or reading texts as hyperlinks and opened in small pop-up windows	multiple
S25		x	O	Mobile tools: WhatsApp, Nearpod, Quizlet and Google Apps	multiple
S26	x		D	TuinLECweb: an intelligent tutoring system that teaches monitoring and self-regulation strategies	formal
S27	x		D	Flip2Learn system	formal
S28		x	D	Wikis	multiple
S29	x		D	multiple tools: an online learning system	formal
S30		x	M	multiple tools: a learning management system and other web 2.0 technologies	multiple
S31		x	M	multiple tools: voice recorder, e-portfolio, colour cards, visual dictionaries, google translator	multiple
S32		x	O	YouTube videos	multiple
S33	x		D	biofeedback	informal
S34	x		O	a mobile self-regulated learning system	formal

Note. x = Yes; S = self-developed; T = third party; M= multiple devices; O= mobile devices; D= desktop PCs.

As 10 studies did not explicitly define the features of Web 2.0 technologies or mobile applications to support SRL, only 24 studies could be further analysed by focusing on how they were designed to facilitate students' SRL processes based on the three phases of SRL (Zimmerman, 2002).

As shown in Figure 4, 21 of the 24 studies adopted technologies that focused mainly on supporting part of the SRL process, such as by providing monitoring affordances in the performance phase, where students could review their learning status. Eight studies highlighted the setting of goals. Ten studies assisted self-regulated

learners in self-reflection. However, only four studies supported the whole process of SRL (Chen et al., 2014; Saks & Leijen, 2019; Shyr & Chen, 2018; Zheng et al., 2018). Chen et al. (2014) adopted an SRL mechanism with a DRAS to support students' reading. The students set learning goals via a self-monitoring table, and they monitored their performance using radar plots. Students could make a self-evaluation. In Saks and Leijen (2019), prompts were added to the learning assignments to assist students' SRL. In addition, Shyr and Chen (2018) adopted the Flip2Learn system to facilitate university students' vocabulary learning and enhance their self-regulatory skills. Furthermore, Zheng et al. (2018) developed a mobile SRL system to assist university students' reading by helping them set goals, make plans, monitor their learning processes and self-evaluate.



4. Discussion

This systematic review provides a synthesis of key findings on the current research status about technology-enhanced SRL from 2011 to 2020 in terms of (1) study characteristics in terms of their publication years and learner types, (2) research methods used to evaluate SRL effectiveness, and (3) the role of technology in SRL.

4.1. Study characteristics

This review indicates that publications on technology-assisted SRL generally increased during the 10-year period. Among the 34 reviewed papers, only seven were published from 2011 to 2015; the rest were published after 2015. Nonetheless, the number of publications declined between 2018 and 2020. Moreover, over 60% of the studies were conducted in tertiary education contexts, followed by secondary school and primary school contexts. These results echo previous findings that many studies on technology-enhanced language learning were conducted in universities (Chang & Hung, 2019; Broadbent & Poon, 2015). Papers targeting kindergarten students are rare. This is probably due to the limited metacognitive abilities of young learners (Alvi & Gillies, 2021; Marulis et al., 2020). Furthermore, some studies (e.g., Bohlmann et al., 2015; Pahiriray, 2021) have revealed that young learners' capability is related to their language ability, which is still under development at their age. Nevertheless, some empirical studies have indicated that preschool children already begin developing an ability for SRL (Lockl & Schneider, 2002; Dignath et al., 2008). The effects of SRL training during development among young learners should be examined. Therefore, future technology-enhanced SRL research can focus more on younger learners, particularly those younger than 9 years.

Additionally, no study on this topic has been conducted on workplace adult learners. The concept of lifelong learning is receiving increasing attention. The results of this study suggest the need to determine how to help workplace adult learners develop SRL skills with technology.

4.2. Research methods

Over half of the studies ($n = 18$) mainly adopted quantitative research methods, and the majority lasted less than four months. As for SRL evaluation, 18 studies analysed both language and SRL outcomes, six studies (17.6%) assessed self-regulation only and 10 studies (29.4%) explored student technology-enhanced SRL profiles using non-experimental research designs. Technology-enhanced SRL had a generally positive effect on language learning outcomes, affective/psychological learning outcomes and students' SRL. For language learning outcomes, quizzes were used in most of the studies. Regarding SRL, the measurements heavily relied on self-report data. However, self-report instruments, such as questionnaires and interviews, were usually deployed

before and/or after the treatment; therefore, students might have overestimated their responses (Roth et al., 2016). Although self-report instruments can reveal students' attitudes and feelings, they can be biased, considering that they depend on how learners perceive themselves. As argued by Greene and Schunk (2017), self-report instruments capture students' perceptions of self-regulation but fail to understand how learners change or adapt self-regulation processes while engaging in learning.

Studies on learners' technology-enhanced language learning (e.g., Lai & Gu, 2011; Shyr & Chen 2018; Zheng et al., 2016) have particularly highlighted factors contributing to technology-enhanced EFL learners' SRL. However, none of these studies examined students' SRL behaviours and the relationship between these behaviours and students' language learning outcomes. Current research on technology-enhanced SRL pays little attention to specific SRL behaviours or strategies of individual learners (Li et al., 2020), that is, learning patterns that share characteristics with SRL behaviours. Only six out of the 34 studies traced students' specific behaviours as indicators of SRL. Some studies have acknowledged the value of using log data derived from technology-enhanced environments as SRL indicators (Araka et al., 2020; Azevedo et al., 2018; Panadero et al., 2016; Winne et al., 2019). Thus, future research can apply mixed research methods to elucidate the characteristics of students' SRL behaviours through longitudinal studies, thereby enriching student perceived SRL with real-time behaviour log data.

Four studies involved teachers' observation (Ferreira et al., 2017; Ghufon & Nurdianingsih, 2019; Karami et al., 2019; Torres et al., 2020). The role of teachers in learning design and interpretation of learning analysis drawn from log data is drawing growing interest (McKenney & Mor, 2015; Persico & Pozzi, 2015; Wen & Song, 2020). Researchers state that teachers should be empowered with necessary analytics knowledge to ensure evidence-based learning support (Ndukwe & Daniel, 2020). It would be interesting to understand language teachers' professional development in teacher inquiry and learning analytics. Such an understanding, along with findings on the characteristics of students' SRL behaviours obtained using log data, would be useful in designing and deploying SRL environments.

4.3. Role of technology

Desktop PCs were the primary devices adopted in the reviewed studies. However, the use of mobile devices increased from 2015 (e.g., Hwang & Fu, 2019; Lin & Lin, 2019). This may be due to the increasing popularity of the use of mobile devices in education in recent years. Because of such proliferation of mobile devices, learning is no longer limited to specific contexts. However, only three studies (8.8%) investigated students' SRL outside the classroom. Lai (2017) suggested that successful language learners often attribute their success to active engagement with the target language beyond the classroom. The findings of this study indicate that further studies can be proposed to explore self-initiated learning activities beyond the classroom and means of supporting learners' SRL with mobile technologies in the future.

In terms of the role of technology in supporting SRL, only four out of the 24 studies examined the entire process of SRL (Chen et al., 2014; Saks & Leijen, 2019; Shyr & Chen, 2018; Zheng et al., 2018). Many models theorise SRL processes, but they share a common understanding that the regulation process should be cyclical and that different phases can influence one another (Panadero, 2017; Zimmerman, 2000). Researchers can design a more systematic tool for supporting all phases of SRL in the context of language learning.

5. Conclusion

The findings of this review identify critical research gaps and have implications for future studies on technology-enhanced SRL. First, this paper presents a systematic review from an analysis of 34 studies published from 2011 to 2020 that focused on investigating SRL in technology-enhanced learning environments. Most of these studies were conducted in tertiary education contexts. Thus, future research may target younger learners, particularly those below the age of 9 years. In addition, the majority of these studies were conducted among undergraduate students. Little is known about postgraduate students' SRL in technology-assisted learning environments. Second, this study sheds light on capturing log data to understand the dynamic nature of SRL and develop technology to support the whole process of SRL. Log data can trace individual events in sequence but cannot explain why learners act in the observed ways and how they characterise their cognitive and metacognitive strategies (Bernacki, 2017). Therefore, comprehensive measurements are needed to understand students' SRL in future research. Prospective studies should utilise technologies to assist students' entire SRL and examine their SRL behaviours. Third, the findings show that technology has been adopted to support the

performance phase of students' SRL more than the two other phases (forethought and self-reflection). More attention should be given to examining students' SRL outcomes than their SRL behaviours.

This review has several limitations. First, the review was not exhaustive; only English-language papers were selected, and data were obtained from only four databases. We will include the Scopus database in our future systematic literature review. Second, in view of the exclusion rate in the study selection in this review, we will specify the subject domains (e.g., language learning, education, and technology) and set the article types (e.g., peer-reviewed articles, workshop papers, and conference papers) when searching databases in the first stage, which may help lower the exclusion rate. Finally, our coding scheme may not be the only possible approach to addressing the research questions. It is therefore suggested that more comprehensive ways of reviewing studies should be investigated and applied in future studies.

Acknowledgement

The study was funded by the General Research Fund (Ref. 18611019), Research Grants Council, University Grant Committee, Hong Kong.

References

- Al Fadda, H. (2019). The relationship between self-regulations and online learning in an ESL blended learning context. *English Language Teaching*, 12(6), 87-93.
- Alvi, E., & Gillies, R. M. (2021). Self-regulated learning (SRL) perspectives and strategies of Australian primary school students: A Qualitative exploration at different year levels. *Educational Review*, 1-23. <https://doi.org/10.1080/00131911.2021.1948390>
- Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—Systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning*, 15(1), 6. <https://doi.org/10.1186/s41039-020-00129-5>
- Ardasheva, Y., Wang, Z., Adesope, O. O., & Valentine, J. C. (2017). Exploring effectiveness and moderators of language learning strategy instruction on second language and self-regulated learning outcomes. *Review of Educational Research*, 87(3), 544-582.
- Azevedo, R., Taub, M., & Mudrick, N. (2018). Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In *Handbook of self-regulation of learning and performance* (pp. 254-270). Routledge/Taylor & Francis Group.
- Bano, M., Zowghi, D., Kearney, M., Schuck, S., & Aubusson, P. (2018). Mobile learning for science and mathematics school education: A Systematic review of empirical evidence. *Computers & Education*, 121, 30-58.
- Ben-Yehudah, G., & Brann, A. (2019). Pay attention to digital text: The impact of the media on text comprehension and self-monitoring in higher-education students with ADHD. *Research in developmental disabilities*, 89, 120-129.
- Bernacki, M. L. (2017). Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: Trace data and their analysis. In *Handbook of self-regulation of learning and performance* (pp. 370-387). Routledge/Taylor & Francis Group.
- Bohlmann, N. L., Maier, M. F., & Palacios, N. (2015). Bidirectionality in self-regulation and expressive vocabulary: Comparisons between monolingual and dual language learners in preschool. *Child Development*, 86(4), 1094-1111.
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1-13.
- Çelik, S., Arkin, E., & Sabriler, D. (2012). EFL learners' use of ICT for self-regulated learning. *Journal of Language and Linguistic Studies*, 8(2), 98-118.
- Chang, M.-M., & Hung, H.-T. (2019). Effects of technology-enhanced language learning on second language acquisition. *Educational Technology & Society*, 22(4), 1-17.
- Chien, C.-W. (2019). Taiwanese EFL undergraduates' self-regulated learning with and without technology. *Innovation in Language Learning and Teaching*, 13(1), 1-16.
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A Meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(3), 231-264.

- Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively?: A Meta-analysis on self-regulation training programmes. *Educational Research Review*, 3(2), 101-129.
- Ducasse, A. M., & Hill, K. (2019). Developing student feedback literacy using educational technology and the reflective feedback conversation. *Practitioner Research in Higher Education*, 12(1), 24-37.
- Ferreira, P. C., Veiga Simão, A. M., & Lopes da Silva, A. (2017). How and with what accuracy do children report self-regulated learning in contemporary EFL instructional settings? *European Journal of Psychology of Education*, 32(4), 589-615.
- Ghufron, M. A., & Nurdianingsih, F. (2019). Flipped teaching with Call in EFL writing class: How does it work and affect learner autonomy? *European Journal of Educational Research*, 8(4), 983-997.
- Greene, J., & Schunk, D. (2017). Historical, contemporary, and future perspectives on self-regulated learning and performance. In *Handbook of self-regulation of learning and performance* (pp. 17-32). Routledge/Taylor & Francis Group.
- Hromalik, C. D., & Koszalka, T. A. (2018). Self-regulation of the use of digital resources in an online language learning course improves learning outcomes. *Distance Education*, 39(4), 528-547.
- Hughes, M. D., Regan, K. S., & Evmenova, A. (2019). A Computer-based graphic organizer with embedded self-regulated learning strategies to support student writing. *Intervention in School and Clinic*, 55(1), 13-22.
- Hung, H.-T., Yang, J. C., Hwang, G.-J., Chu, H.-C., & Wang, C.-C. (2018). A Scoping review of research on digital game-based language learning. *Computers & Education*, 126, 89-104.
- Hwang, G.-J., & Fu, Q.-K. (2019). Trends in the research design and application of mobile language learning: A Review of 2007–2016 publications in selected SSCI journals. *Interactive Learning Environments*, 27(4), 567-581.
- Lai, C. (2017). *Autonomous language learning with technology: Beyond the classroom*. Bloomsbury Publishing.
- Lai, C., & Gu, M. Y. (2011). Self-regulated out-of-class language learning with technology. *Computer Assisted Language Learning*, 24(4), 317-335.
- Lee, S.-M. (2019). A Systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning*, 25(3), 294-318. <https://doi.org/10.1080/09588221.2019.1688836>
- Lehmann, T., Hähnlein, I., & Ifenthaler, D. (2014). Cognitive, metacognitive and motivational perspectives on preflexion in self-regulated online learning. *Computers in Human Behavior*, 32, 313-323.
- Li, S., Chen, G., Xing, W., Zheng, J., & Xie, C. (2020). Longitudinal clustering of students' self-regulated learning behaviors in engineering design. *Computers & Education*, 153, 103899. <https://doi.org/10.1016/j.compedu.2020.103899>
- Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A Systematic review and meta-analysis. *Computer Assisted Language Learning*, 32(8), 878-919. <https://doi.org/10.1080/09588221.2018.1541359>
- Lin, C.-C., Lin, V., Liu, G.-Z., Kou, X., Kulikova, A., & Lin, W. (2019). Mobile-assisted reading development: A Review from the Activity Theory perspective. *Computer Assisted Language Learning*, 33(8), 833-864.
- Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, 26(4), 327-333.
- Marulis, L. M., Baker, S. T., & Whitebread, D. (2020). Integrating metacognition and executive function to enhance young children's perception of and agency in their learning. *Early Childhood Research Quarterly*, 50, 46-54.
- McKenney, S., & Mor, Y. (2015). Supporting teachers in data-informed educational design. *British Journal of Educational Technology*, 46(2), 265-279.
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*, 4(1), 1. <https://doi.org/10.1186/2046-4053-4-1>
- Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics, value and tools for teacher data literacy: A Systematic and tripartite approach. *International Journal of Educational Technology in Higher Education*, 17(1), 1-31.
- Panadero, E. (2017). A Review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Panadero, E., Klug, J., & Järvelä, S. (2016). Third wave of measurement in the self-regulated learning field: When measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research*, 60(6), 723-735.
- Pahuriray, V. G. M. (2021). Self-regulating capacity in language learning and English academic achievement. *Globus Journal of Progressive Education*, 11(2), 82-86.
- Palalas, A., & Wark, N. (2020). The relationship between mobile learning and self-regulated learning: A Systematic review. *Australasian Journal of Educational Technology*, 36(4), 151-172.

- Paris, S. G., & Newman, R. S. (1990). Developmental aspects of self-regulated learning. *Educational Psychologist*, 25(1), 87-102.
- Paris, S. G., & Winograd, P. (1999). The role of self-regulated learning in contextual teaching: Principles and practices for teacher preparation. In *Contextual teaching and learning: Preparing teachers to enhance student success in and beyond school* (pp. 219-252). ERIC Clearinghouse on Teaching and Teacher Education, AACTE.
- Persico, D., & Pozzi, F. (2015). Informing learning design with learning analytics to improve teacher inquiry. *British Journal of Educational Technology*, 46(2), 230-248.
- Pintrich, P. R. (2000). The Role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451-502). Academic Press.
- Roth, A., Ogrin, S., & Schmitz, B. (2016). Assessing self-regulated learning in higher education: A Systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*, 28(3), 225-250.
- Saks, K., & Leijen, Ä. (2019). The Efficiency of prompts when supporting learner use of cognitive and metacognitive strategies. *Computer Assisted Language Learning*, 32(1-2), 1-16.
- Shadiev, R., & Yang, M. (2020). Review of studies on technology-enhanced language learning and teaching. *Sustainability*, 12(2), 524. <https://doi.org/10.3390/su12020524>
- Sharples, M., de Rooock, R., Ferguson, R., Gaved, M., Herodotou, C., Koh, E., Kukulska-Hulme, A., Looi, C.-K., McAndrew, P., Rienties, B., Weller, M., & Wong, L. H. (2016). *Innovating pedagogy 2016: Open University innovation report 5*. Institute of Educational Technology, The Open University.
- Shih, K.-P., Chen, H.-C., Chang, C.-Y., & Kao, T.-C. (2010). The Development and implementation of scaffolding-based self-regulated learning system for e/m-learning. *Educational Technology & Society*, 13(1), 80-93.
- Shyr, W. J., & Chen, C. H. (2018). Designing a technology-enhanced flipped learning system to facilitate students' self-regulation and performance. *Journal of Computer Assisted Learning*, 34(1), 53-62.
- Stemler, S. E. (2004). A Comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, 9(1), 4. <https://doi.org/10.7275/96jp-xz07>
- Su, Y., Zheng, C., Liang, J.-C., & Tsai, C.-C. (2018). Examining the relationship between English language learners' online self-regulation and their self-efficacy. *Australasian Journal of Educational Technology*, 34(3). <https://doi.org/10.14742/ajet.3548>
- Tao, J., Zheng, C., Lu, Z., Liang, J.-C., & Tsai, C.-C. (2020). Cluster analysis on Chinese university students' conceptions of English language learning and their online self-regulation. *Australasian Journal of Educational Technology*, 36(2), 105-119.
- Torres, M. C. C., Salamanca, Y. N. S., Cely, J. P. C., & Aguilar, J. L. B. (2020). All we need is a boost! Using multimodal tools and the translanguaging strategy: Strengthening speaking in the EFL classroom. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 10(3), 28-47.
- Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: A Review of empirical research. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 524-533). <https://doi.org/10.1145/3375462.3375483>
- van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology*, 31(5), 508-519.
- Winne, P. H., Teng, K., Chang, D., Lin, M. P.-C., Marzouk, Z., Nesbit, J. C., Patzak, A., Raković, M., Samadi, D., & Vytasek, J. (2019). nStudy: Software for learning analytics about processes for self-regulated learning. *Journal of Learning Analytics*, 6(2), 95-106.
- Woottipong, K. (2022). Facilitating learners' self-regulated learning skills and self-efficacy to write in English using technologies. *Acuity: Journal of English Language Pedagogy, Literature and Culture*, 7(1), 101-122.
- Yang, T. C., Chen, M. C., & Chen, S. Y. (2018). The Influences of self-regulated learning support and prior knowledge on improving learning performance. *Computers & Education*, 126, 37-52.
- Yeh, Y. C., Kwok, O. M., Chien, H. Y., Sweany, N. W., Baek, E., & McIntosh, W. A. (2019). How college students' achievement goal orientations predict their expected online learning outcome: The mediation roles of self-regulated learning strategies and supportive online learning behaviors. *Online Learning*, 23(4), 23-41.
- Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The Impact of gamification on learning and instruction: A Systematic review of empirical evidence. *Educational Research Review*, 30, 100326.
- Zhai, X., Fang, Q., Dong, Y., Wei, Z., Yuan, J., Cacciolatti, L., & Yang, Y. (2018). The Effects of biofeedback-based stimulated recall on self-regulated online learning: A Gender and cognitive taxonomy perspective. *Journal of Computer Assisted Learning*, 34(6), 775-786.

- Zheng, L., Li, X., & Chen, F. (2018). Effects of a mobile self-regulated learning approach on students' learning achievements and self-regulated learning skills. *Innovations in Education and Teaching International*, 55(6), 616-624.
- Zhou, Y., & Wei, M. (2018). Strategies in technology-enhanced language learning. *Studies in Second Language Learning and Teaching*, 8(2), 471-495.
- Zou, D., Huang, Y., & Xie, H. (2019). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, 1-27.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice*, 41(2), 64-70.

Appendix A. List of selected studies

-
- | | |
|-----|---|
| S1 | Ahn, T. Y., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. <i>British Journal of Educational Technology</i> , 47(4), 778-786. |
| S2 | Al Fadda, H. (2019). The relationship between self-regulations and online learning in an ESL blended learning context. <i>English Language Teaching</i> , 12(6), 87-93. |
| S3 | Alzubi, A. A. F., & Singh, M. K. A. P. M. (2018). The impact of social strategies through smartphones on the Saudi learners' socio-cultural autonomy in EFL reading context. <i>International Electronic Journal of Elementary Education</i> , 11(1), 31-40. |
| S4 | Çelik, S., Arkin, E., & Sabriler, D. (2012). EFL learners' use of ICT for self-regulated learning. <i>Journal of Language and Linguistic Studies</i> , 8(2), 98-118. |
| S5 | Chen, C.-M., Chen, L.-C., & Yang, S.-M. (2019). An English vocabulary learning app with self-regulated learning mechanism to improve learning performance and motivation. <i>Computer Assisted Language Learning</i> , 32(3), 237-260. |
| S6 | Chen, C.-M., Wang, J.-Y., & Chen, Y.-C. (2014). Facilitating English-language reading performance by a digital reading annotation system with self-regulated learning mechanisms. <i>Educational Technology & Society</i> , 17(1), 102-114. |
| S7 | Chen, Y.-L., & Hsu, C.-C. (2020). Self-regulated mobile game-based English learning in a virtual reality environment. <i>Computers & Education</i> , 103910. https://doi.org/10.1016/j.compedu.2020.103910 |
| S8 | Chen, & Lee, S. Y. (2018). Application-driven educational game to assist young children in learning English vocabulary. <i>Educational Technology & Society</i> , 21(1), 70-81. |
| S9 | Chien, C.-W. (2019). Taiwanese EFL undergraduates' self-regulated learning with and without technology. <i>Innovation in Language Learning and Teaching</i> , 13(1), 1-16. |
| S10 | Ferreira, P. C., Veiga Simão, A. M., & Lopes da Silva, A. (2017). How and with what accuracy do children report self-regulated learning in contemporary EFL instructional settings? <i>European Journal of Psychology of Education</i> , 32(4), 589-615. |
| S11 | Ghufron, M. A., & Nurdianingsih, F. (2019). Flipped teaching with Call in EFL writing class: How does it work and affect learner autonomy? <i>European Journal of Educational Research</i> , 8(4), 983-997. |
| S12 | Hong, J. C., Hwang, M. Y., Chang, H. W., Tai, K. H., Kuo, Y. C., & Tsai, Y. H. (2015). Internet cognitive failure and fatigue relevant to learners' self-regulation and learning progress in English vocabulary with a calibration scheme. <i>Journal of Computer Assisted Learning</i> , 31(5), 450-461. |
| S13 | Hromalik, C. D., & Koszalka, T. A. (2018). Self-regulation of the use of digital resources in an online language learning course improves learning outcomes. <i>Distance Education</i> , 39(4), 528-547. |
| S14 | Karami, S., Sadighi, F., Bagheri, M. S., & Riasati, M. J. (2019). The impact of application of electronic portfolio on undergraduate English majors' writing proficiency and their self-regulated learning. <i>International Journal of Instruction</i> , 12(1), 1319-1334. |
| S15 | Kondo, M., Ishikawa, Y., Smith, C., Sakamoto, K., Shimomura, H., & Wada, N. (2012). Mobile assisted language learning in university EFL courses in Japan: Developing attitudes and skills for self-regulated learning. <i>ReCALL</i> , 24(2), 169-187. |
| S16 | Lai, C., & Gu, M. Y. (2011). Self-regulated out-of-class language learning with technology. <i>Computer Assisted Language Learning</i> , 24(4), 317-335. |
| S17 | Lei, Z. (2018). Vocabulary learning assisted with smart phone application. <i>Theory and Practice in Language Studies</i> , 8(11), 1511-1516. |
| S18 | Lin, V., Liu, G. Z., & Chen, N. S. (2020). The effects of an augmented-reality ubiquitous writing application: A comparative pilot project for enhancing EFL writing instruction. <i>Computer Assisted Language Learning</i> , 1-42. |
| S19 | Liu, S. H. J., Lan, Y. J., & Ho, C. Y. Y. (2014). Exploring the relationship between self-regulated vocabulary learning and web-based collaboration. <i>Educational Technology & Society</i> , 17(4), 404-419. |
| S20 | Llorens, A., Vidal-Abarca, E., & Cerdán, R. (2016). Formative feedback to transfer self-regulation of task-oriented reading strategies. <i>Journal of Computer Assisted Learning</i> , 32(4), 314-331. |
| S21 | Man-Kit, L. E. E., & Evans, M. (2019). Investigating the operating mechanisms of the sources of L2 writing self-efficacy at the stages of giving and receiving peer feedback. <i>The Modern Language Journal</i> , 103(4), 831-847. |
| S22 | Naseri, S., & Motallebzadeh, K. (2016). Podcasts: A Factor to improve Iranian EFL learner's self-regulation ability and use of technology. <i>Educational Technology & Society</i> , 19(2), 328-339. |
| S23 | Roussel, S. (2011). A computer assisted method to track listening strategies in second language learning. <i>ReCALL</i> , 23(2), 98-116. |
-

-
- S24** Saks, K., & Leijen, Ä. (2019). The efficiency of prompts when supporting learner use of cognitive and metacognitive strategies. *Computer Assisted Language Learning*, 32(1-2), 1-16.
- S25** Seifert, T., & Har-Paz, C. (2020). The effects of mobile Learning in an EFL class on self-regulated learning and school achievement. *International Journal of Mobile and Blended Learning (IJMBL)*, 12(3), 49-65.
- S26** Serrano, M. Á., Vidal-Abarca, E., & Ferrer, A. (2018). Teaching self-regulation strategies via an intelligent tutoring system (TuinLECweb): Effects for low-skilled comprehenders. *Journal of Computer Assisted Learning*, 34(5), 515-525.
- S27** Shyr, W. J., & Chen, C. H. (2018). Designing a technology-enhanced flipped learning system to facilitate students' self-regulation and performance. *Journal of Computer Assisted Learning*, 34(1), 53-62.
- S28** Su, Y., Li, Y., Liang, J.-C., & Tsai, C.-C. (2019). Moving literature circles into wiki-based environment: The Role of online self-regulation in EFL learners' attitude toward collaborative learning. *Computer Assisted Language Learning*, 32(5-6), 556-586.
- S29** Su, Y., Zheng, C., Liang, J.-C., & Tsai, C.-C. (2018). Examining the relationship between English language learners' online self-regulation and their self-efficacy. *Australasian Journal of Educational Technology*, 34(3).
- S30** Tao, J., Zheng, C., Lu, Z., Liang, J.-C., & Tsai, C.-C. (2020). Cluster analysis on Chinese university students' conceptions of English language learning and their online self-regulation. *Australasian Journal of Educational Technology*, 36(2), 105-119.
- S31** Torres, M. C. C., Salamanca, Y. N. S., Cely, J. P. C., & Aguilar, J. L. B. (2020). All we need is a boost! Using multimodal tools and the translanguaging strategy: Strengthening speaking in the EFL classroom. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 10(3), 28-47.
- S32** Wang, H.-C., & Chen, C. W.-y. (2020). Learning English from YouTubers: English L2 learners' self-regulated language learning on YouTube. *Innovation in Language Learning and Teaching*, 14(4), 333-346.
- S33** Zhai, X., Fang, Q., Dong, Y., Wei, Z., Yuan, J., Cacciolatti, L., & Yang, Y. (2018). The effects of biofeedback-based stimulated recall on self-regulated online learning: A Gender and cognitive taxonomy perspective. *Journal of Computer Assisted Learning*, 34(6), 775-786.
- S34** Zheng, L., Li, X., & Chen, F. (2018). Effects of a mobile self-regulated learning approach on students' learning achievements and self-regulated learning skills. *Innovations in Education and Teaching International*, 55(6), 616-624.
-

Exploring the Research Trajectory of Digital Game-based Learning: A Citation Network Analysis

Wiwit Ratnasari¹, Tzu-Chuan Chou^{1*} and Chen-Hao Huang²

¹Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C. // ²Graduate Institute of Technology Management, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C. // ratnasariwiwit@gmail.com // tcchou@mail.ntust.edu.tw // chhuang@mail.ntust.edu.tw

*Corresponding author

(Submitted November 9, 2022; Revised April 2, 2022; Accepted May 9, 2022)

ABSTRACT: The digital revolution has heavily influenced digital game-based learning, yet as the revolution progresses, the conception of such learning has shifted along with the increasing complexity of the digital environment. Our study thus aims to identify research standing at this important juncture and to explain the shift in digital game-based learning research fields by adopting an integrated approach of main path analysis that yields this topic's knowledge diffusion. Using key-route 8 to construct the path, we collect a total of 2156 articles and their data from The Web of Science database. From over 30 years of digital game-based learning development, 26 of the most influential studies are identified and visualized using Pajek software. The findings show two development phases for this field: exploring the role of gaming for educational purpose as well as facilitating learning performance. The research focus in the first phase prominently explores the potentials of digital games for educational purposes, and then the focus evolves in the second phase into actualizing the identified potentials. We propose a framework of digital game-based learning affordance actualization to explain these shifting phenomena in the specific research fields. Furthermore, unveiling the changing conception of digital game-based learning research is important for instructional designers, scholars, and educators to truly understand how technology can enhance teaching and facilitate learning performance.

Keywords: Affordance actualization, Main path analysis, Digital game-based learning, Citation network, Knowledge diffusion

1. Introduction

Research on game-based learning has attracted widespread scholarly attention over the last few decades. Due to the rapid growth and public acceptance of technologies, academic studies related to digital game-based learning (DGBL) have grown very fast since 2006 (Hwang & Wu, 2012). Although the digital revolution has heavily influenced DGBL, its concept is not only growing alongside with the development of digital technologies, but also changing the education paradigm. The integration of modern technology and gaming in the learning and educational context is not a new concept, yet the broad spread of digital game acceptance has attracted instructional designers, researchers, and educators to further explore its potential (Plass et al., 2015). DGBL started out by focusing on the usage of digital technology to increase students' learning achievement, but over the years it shifted its major attention to the alignment between technology and learning environment to satisfy numerous learning needs. Over the last three decades, influenced by technology advancement and entertainment gaming trends, the concept of DGBL has been constantly changing.

Over this course of time, an abundant amount of studies has proven the effectiveness of educational computer games to support learning programs as a way for increasing student motivation and engagement in various subject areas, such as natural science courses (Hwang et al., 2013; Sung & Hwang, 2013), English as a foreign language (Huang & Huang, 2015; Lin et al., 2020), mathematics (Ke, 2008; Ku et al., 2014), computer science and engineering (Coller & Scott, 2009; Ebner & Holzinger, 2007), health (Quail & Boyle, 2019), and geography (Tüzün et al., 2009). Moreover, with the number of publications on this subject growing, several review studies have been conducted to identify the current development and research trends of the DGBL field over a certain period (Hwang & Wu, 2012; Tsai & Fan, 2013). Cheng et al. (2020) show the importance of reviewing based on highly cited articles, arguing that such articles represent highly valued topics and important trends due to having solid pedagogical theories and well-recognized data analysis methods.

Liu and Lu (2012) conversely point out methods that identify the most significant path in a large citation network, which is main path analysis. In contrast to citation counts, main path analysis not only considers the direct influences, but also takes indirect influences into account. Assuming that citation links represent diffusion

of knowledge from one work to another, this method assigns values to the link that connects two documents instead of directly assigning values to the documents themselves. Therefore, this method identifies significant “links” in lieu of important “nodes.” The nodes associated with the link are still considered important. Thus, the results obtained from citation counts might be different from the results obtained from the main path method. This method is also effective in highlighting a sequence of major historical development events in a complex citation network. Therefore, main path analysis is a powerful strategy for tracing the evolution of a science or technology throughout history (Liu et al., 2019; Liu & Lu, 2012).

Regardless of the valuable insights that have been provided from previous review studies, large and complex citation networks have continuously emerged as a side effect of the rapidly increasing interest in this topic every year. To the best of our knowledge, no papers have employed a study of significant historical development events in a complex citation network of DGBL. Analyzing the citation network of DGBL can help us identify the critical intellectual development milestones of DGBL studies and identify their development trajectory. Understanding the exponential growth of the DGBL literature can also reveal the dynamic nature of this field. Given this importance, surprisingly little if any research has been conducted on the origin and how DGBL studies evolve over time.

The trajectory of knowledge, resulting from main path visualization, can tell us something about the changing nature of the digital technology and learning nexus that needs to be explained. Furthermore, actualizing the potential of a digital game in education, represented by affordance, may contribute to the success of DGBL and improve the learning experience. All in all, conducting main path analysis to find the most important works in the DGBL research fields and explaining how these fields have evolved are imperative and beneficial.

The purpose of this study is to identify works standing at an important juncture and to explain the shift in DGBL research fields through main path analysis. This study utilizes the Web of Science Database covering the Science Citation Index Expanded (SCI-EXPANDED) and Social Sciences Citation Index (SSCI) about the subject of DGBL. The research questions are listed as follows.

- RQ1.** Which research studies have the most influence on the development of the digital game-based learning literature?
- RQ2.** How has the research of digital game-based learning evolved?

To answer these research questions, this study adopts main path analysis and utilizes systematic literature research to trace 30 years of DGBL’s development trajectory. Furthermore, Strong et al. (2014) affordance actualization lens is applied as a theoretical method to further understand the shifting phenomena in the DGBL research field (Hwang et al., 2021).

This study fills the gap in the literature and contributes to DGBL research by proposing several distinctions. First, in contrast to previous review papers focusing on a certain aspect of DGBL or based on highly cited articles, this study targets all research that has been published in the 30 years of DGBL. Second, we identify the most significant works based on the citation network of DGBL research fields. Third, knowing the most important juncture in the historical development of DGBL research fields will tell us something about the changing nature in their evolution. Thus, in this study we shall explain the evolution of DGBL research fields.

2. Research methodology

2.1. Data collection

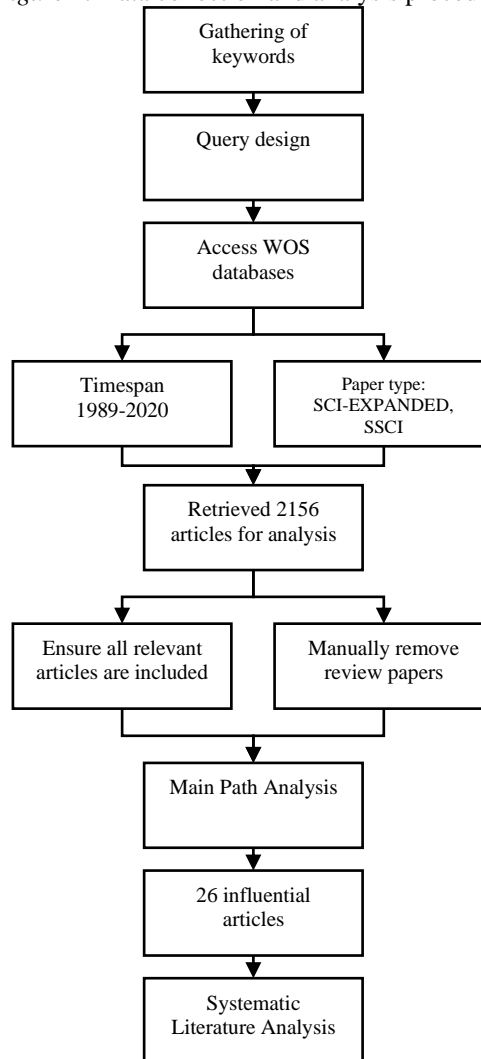
Figure 1 below explains the procedure of data collection and evaluation. Our goal is to have a complete dataset, which includes as many relevant articles as possible and excludes those that are irrelevant. To achieve its goal, this study follows four steps of data collection by Chuang et al. (2017).

First, we list several related keywords by reviewing five recent review articles in the field of DGBL (Acquah & Katz, 2020; Hung et al., 2020; Lai & Bower, 2020; Martin et al., 2020; Noroozi et al., 2020). Second, the keywords gathered help further design our search strategy to narrow the search results. Based on the searching strategy, the following query is used to narrow the search result:

TS=((game-based AND learn) OR (gamif* AND learn*) OR "education* computer game*" OR "serious computer game*" OR "digital serious game*") NOT TI=review*

The Web of Science core collection is used to search and collect published articles that are going to be used for our study. Our datasets cover the Science Citation Index Expanded (SCI-EXPANDED) and Social Sciences Citation Index (SSCI) from 1991 until the day of data collection, which is May 18, 2021. In the third step, we aim to ensure all relevant articles are included in our dataset by reviewing reference sections of selected review papers mentioned previously and manually adding any missing papers into our dataset. Furthermore, we make sure that all the highly cited DGBL papers in the WOS Database are included in our dataset. To get an objective result, we purposely exclude review papers, because in general they are highly cited not because of their original ground-breaking results, but rather due to their comprehensive summary of the results of a field (Ho et al., 2017; Liu & Lu, 2012). Thus, in the last steps we manually remove review papers in the datasets. The search query results in 2156 research articles. We then export all the record content and citation information of the search results and proceed with constructing main path analysis.

Figure 1. Data collection and analysis procedure



2.2. Main path analysis

The total number of all publications in a scientific field is relatively large, and therefore a quantitative approach is needed to make it possible to analyze the large data in a citation network. Main path analysis is a citation-based method that extracts the backbone of a large citation network. It was first introduced by Hummon and Dereian (1989), assuming that knowledge from previous research disperses to later research through citations. Main path analysis calculates the extent to which a particular citation or article is needed for linking articles (Nooy et al., 2018). In that sense, citations that are needed in paths connecting many articles carry more significance than those that are barely linked to any articles (Calero-Medina & Noyons, 2008; Nooy et al., 2018). As the citation network in a scientific field is growing rapidly, the citation network is becoming massive and more complex. Batagelj (2003) advance Hummon and Dereian's (1989) weights by proposing efficient

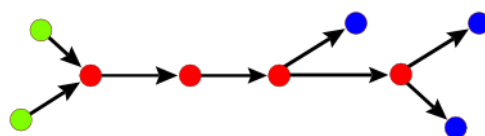
algorithms for determining various versions of the significance index, so that they can be used for analysis of a very large citation network. They implemented the algorithm in Pajek software for analysis of large networks, and thus their advancements accelerated the use of path analysis (Batagelj, 2003; Lathabai et al., 2018; Liu & Lu, 2012)

The original main path approach does have some limitations (Liu & Lu, 2012). In the original main path method, the search procedure is to find the single most significant path for the whole network. As Liu and Lu (2012) note, the path resulting from this approach cannot guarantee that this single path is the most significant among all paths in the whole network, and it does not allow to find the significant nodes that bring together ideas from many earlier publications. To overcome this limitation, they set up an approach that integrates several methods into one analysis, which are the global method, backward method, multiple main path method, and key-route search method.

This breakthrough, particularly the key-route method, is an excellent tool for visually displaying the development structure of an entire scientific field, which suggests a divergence-convergence-divergence process on that structure (Liu et al., 2019; Liu & Lu, 2012). The key-route approach ensures that significant top links in the citation network are included in the main paths, thus complementing the aftermentioned limitations. The path constructed from the key-route method is based on the most significant links with the highest traversal counts as a seed link, and then it searches forward and backward until a “source” and a “sink” are hit. The path is then constructed by connecting all the resulting networks. Although the key-route approach has the possibility to determine as many seed links as possible, the higher the seed link number is that ones decide, the more complex the network will be.

Main path analysis operates in two steps. It first determines the traversal count of each citation link from each source to each sink. It then searches for the main path by linking citation links based on size of traversal counts. A number of terms must be defined to precede with the discussion of traversal weights. As shown in Figure 2, there are 3 types of nodes in a citation network: source, intermediate, and sink. Source nodes are articles that are cited by others, but are not citing within the datasets. Intermediate nodes are articles that are citing and cited by others. Sink nodes are articles that are citing others, but not cited within the datasets. Apart from nodes, in the citation network there are arrows as well. The arrows denote their links; the thickness of the line indicates the traversal counts of the links. Thus, the thicker the line is, the more significant is the link (Liu & Lu, 2012).

Figure 2. A simple citation network



The three basic methods to determine the traversal weight are SPC (search path count), SPLC (search path link count), and SPNP (search path node pair). The method in calculating search path count is determined by how to define source and sink. In the SPC method, traversal weight is calculated from the number of links traversed by all possible paths from all sources (green nodes) to all sinks (blue nodes). In the SPLC method, all nodes before a particular link are also seen as a source, including intermediate nodes (red nodes). In SPNP, all nodes before a particular link are seen as a source, and all nodes after a particular link are seen as a sink.

A key differentiation between SPC and SPNP is that SPC sees the intermediate node as merely an intermediary for knowledge to flow, but SPNP considers it a knowledge depository as well, whereas knowledge diffusion in the scientific and technological world does not work this way. Thus, the most appropriate method for calculating the search path is SPLC. This is because SPLC treats intermediate nodes that are not only seen as passing the knowledge, but also are knowledge sources as well. Therefore, for tracing the knowledge diffusion trajectory in scientific and technological development, Liu et al. (2019) note that SPLC is the most recommended search path count method.

3. Findings and analysis

The trend of DGBL research is still growing. After performing main path analysis from 1991 to May 18, 2021, the total number of works published adds up to 2,156. We apply Loglet Analysis to the publication data to see the growth trend of publications studying this topic. Loglet Analysis is a logistical trend analysis tool designed to

analyze sets of time-series data and decompose the growth process into S-shape logistic components (Meyer et al., 1999).

Figure 3 below shows the growing trend of DGBL research over the years. As seen in the figure, its growing trend has continued to increase with attention and interest over time. Our findings are in line with previous studies demonstrating that DGBL studies have become more and more important over the past decade, as many researchers recognize the potential benefits of computer games for learning (Chen et al., 2020; Hwang & Wu, 2012). Based on the analysis, if this trend persists, then we predict that the growth of this research topic will continue to expand until 2030.

Figure 3. DGBL research growth trend (historical and projected)

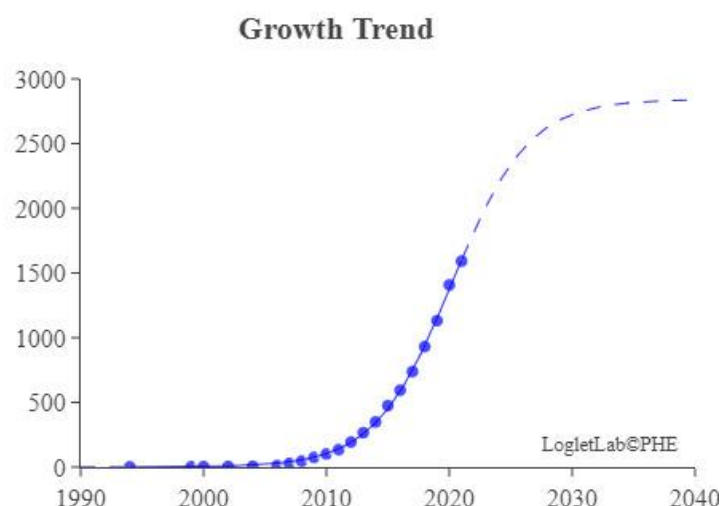
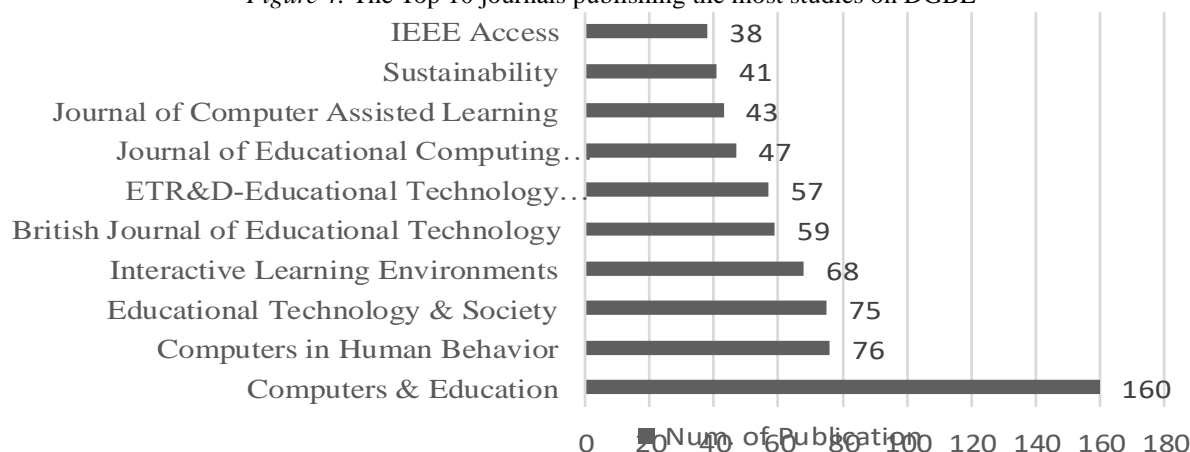


Figure 4 shows the main journals that publish the most studies on DGBL. We count the number of publications in the DGBL field and show only the top 10 list of journals. Computers & Education has the greatest number of publications due to its scope, which focuses on pedagogical uses of digital technology, while Computer in Human Behavior focuses on wider topics that cover the use of computers from a psychological perspective, including the psychological effects of computers on learning. Educational Technology & Society is in the third position in terms of its publication number of DGBL research. Although ET&S and Computer in Human Behavior focus more on research in educational technology, their number of publications does not exhibit much of a difference.

Figure 4. The Top 10 journals publishing the most studies on DGBL



3.1. Cross-citation network of the DGBL literature

The cross-citation network shows how knowledge is created and distributed among authors. Analyzing the cross-citation network is one way to observe authors' correlation with each other. In the knowledge creation process,

3.2. Main path analysis of digital game-based learning

We perform main path analysis to uncover the more critical development path in DGBL research. The main path is represented by using key-route 8, which means the path is constructed based on the top 8 links with the highest traversal counts, and then it searches forward and backward until a “source” and a “sink” are hit. We believe that key-route 8 adequately represents prominent junctures in DGBL research fields. Our observations indicate that numerous significant links are represented, while paths of lesser significance are ignored, thus offering a bird’s eye view of the complex citation network of DGBL.

Table 1 below is the list of the highest traversal links used as a seed link in constructing the network in this analysis. The highest traversal links are believed to be the most significant path, because they bring together ideas from many earlier publications. Thus, the most influential articles that contribute to the development of the DGBL research field can be viewed as the main path constructed from the most significant publications.

Table 1. The highest traversal links

Counts	Traversal Counts (SPLC)	Routes
1	179240.00	HwangSHYH2013 => HwangSHHT2012
2	177632.00	HungSY2015 => HamariSRCAE2016
3	154656.00	SungH2013 => HwangYW2013
4	138804.00	HwangWC2012 => SungH2013
5	132348.00	HwangSHHT2012 => SungH2013
6	86112.00	SungH2013 => HungSY2015
7	85284.00	HwangYW2013 => HwangCC2015
8	84666.00	HwangYW2013 => HwangHC2014

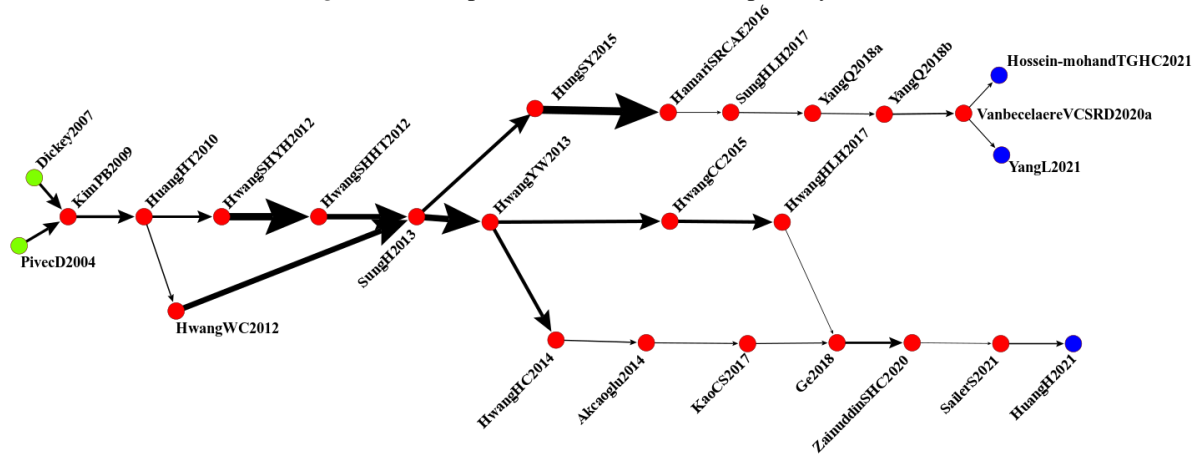
Table 2. Papers by region focus

Label	Author(s)	Region Focus
PivecD2004	Pivec & Dziabenko (2004)	Austria
Dickey2007	Dickey (2007)	United States
KimPB2009	Kim et al. (2009)	South Korea
HuangHT2010	Huang et al. (2010)	United States
HwangWC2012	Hwang et al. (2012c)	Taiwan
HwangSHHT2012	Hwang et al. (2012a)	Taiwan
SungH2013	Sung & Hwang (2013)	Taiwan
HwangYW2013	Hwang et al. (2013)	Taiwan
HwangSHYH2012	Hwang et al. (2012b)	Taiwan
HwangHC2014	Hwang et al. (2014)	Taiwan
Akcaoglu2014	Akcaoglu (2014)	Turkey
HwangCC2015	Hwang et al. (2015)	Taiwan
HungSY2015	Hung et al. (2015)	Taiwan
HamariSRCAE2016	Hamari et al. (2016)	United States
HwangHLH2017	Hwang et al. (2017)	Taiwan
SungHLH2017	Sung et al. (2017b)	Taiwan
KaoCS2017	Kao et al. (2017)	Taiwan
YangQ2018a	Yang & Quadir (2018)	Taiwan
YangLC2018	Yang et al. (2018)	Taiwan
Ge2018	Ge (2018)	China
YangQ2018b	Yang & Quadir (2018)	Taiwan
ZainuddinSHC2020	Zainuddin et al. (2020)	Indonesia
VanbecelaereVCSRD2020b	Vanbecelaere et al. (2020)	Belgium
SailerS2020	Sailer & Sailer (2020)	Germany
HuangH2021	Huang & Hew (2021)	Hong Kong
Hossein-mohandTGHC2021	Hossein-Mohand et al. (2021)	Spain

Figure 6 shows the critical development path of DGBL studies. The finding reveals that there are 26 most influential articles published from 2004 to 2021. The data are then visualized using the Pajek software. The nodes are represented by a different color. Green nodes are the source nodes, red nodes are intermediate nodes, and blue nodes are sink nodes. The arrows denote the direction of knowledge flow, and the thickness of the line indicates the traversal counts of the links. The thicker the line is, the more significant is the link (Liu & Lu, 2012).

Each paper is assigned with a code. This code consists of the last name of the first author, followed by the latter authors' initials, and ends with the publication year. As an example, the study from Pivec and Dziabenko in 2004 is coded as PivecD2004. In case there are duplicate codes, then the codes have lower-case letters added at the end, such as YangQ2018a and YangQ2018b, which are written by Jie Chi Yang and Benazir Quadir in the same year.

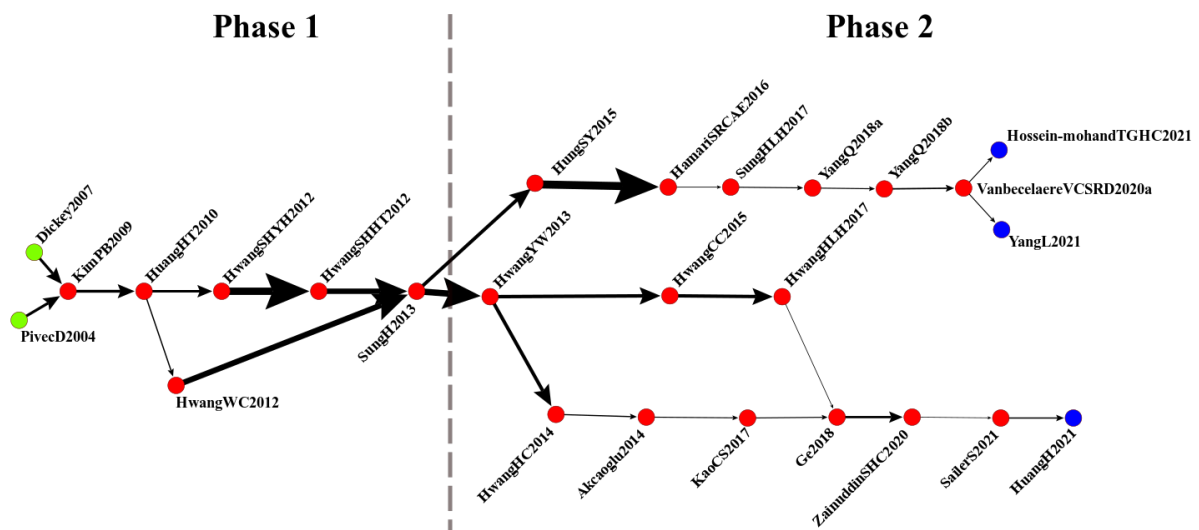
Figure 6. Main path of DGBL studies (Top 8 Key-route)



Based on main path analysis, we find that the most influential articles in the DGBL research field have been carried out in Taiwan, followed by United States (see Table 2). The author who has appeared the most out of the 26 most influential articles is Gwo-Jen Hwang, appearing in 9 out of 26 most influential articles, followed by Jie Chi Yang appearing in 3 articles.

We next study the articles to better understand the main path results. From studying those articles on the main path, we divide the DGBL research development into two major stages (Figure 7). The early days of DGBL research appear to focus on the development of an interactive learning environment and demonstrate how various aspects of computer games have great potential to improve the learning process. In the second phase, the trends of DGBL research bifurcate into three areas.

Figure 7. Development trajectory of digital game-based learning studies



3.2.1. Exploring the role of gaming for educational purposes

Games have evolved from traditional games to handheld electronic games, and along with the rise of the Internet in the 1990s, online gaming started to thrive. Online gaming allows players to link up together, thus enabling them to collaborate and/or compete to win a game. Its popularity has attracted the attention of researchers and educators to explore its potential in providing a new innovative way of learning (Plass et al., 2015). Driven by

the desire to enhance the learning environment by presenting this new and innovative technology to build digital games for educational purposes, DGBL has thus become an emerging research topic.

Two critical sources that shape today's DGBL research are the studies of Maja Pivec and Olga Dziabenko in 2004 and Michele D. Dickey in 2007. Pivec and Dziabenko (2004) propose a new way of learning by introducing a collaborative learning-social skill game concept based on the constructivist learning approach and collaborative learning, UniGame. Their study contributes to how collaborative learning can be applied in a fun and engaging way with the use of (digital single-player) online-role playing games.

Along with the trends of casual games, DGBL research trends also shifted. In the 2000s, a ground-breaking game genre emerged. One of the most popular game genres is the Massively Multiple Online Role-Playing Game (MMORPG), which allows thousands of players to interact at the same time in an online gameplay environment. The popularity of the MMORPG game genre soon attracted instructional designers, researchers, and educators to further explore its potential. Examining how some elements in MMORPG's game design can support intrinsic motivation, Dickey (2007) suggests that its design may provide a practical model for creating interactive learning environments by providing choice, control, collaboration, challenge, and achievement. Her findings give insightful contributions to instructional designers and educators to develop an interactive and engaging learning environment.

Keeping in line with the source knowledge, researchers in the following years continued to explore the role of digital game for educational purposes by developing educational computer game design, features, and game concepts (Tsai & Fan, 2013). Hwang et al. (2012c) redefine the research in the field by developing a competitive board game with an online game approach. Later in the same year, Hwang et al. (2012b) state that without proper learning strategies or supportive models, the learning achievement of students might not be as good as expected. Thus, they propose a knowledge engineering approach for developing educational computer games, the Repertory Grid Method. This approach shows significant improvement for students' learning performance in differentiating knowledge. They also realize that some issues need to be further investigated.

In their next research, Hwang et al. (2012a) investigate the effect of students' learning styles on their performance. They present a personalized educational computer game and examine its effectiveness in improving the learning achievement of students. They argue that students' learning styles are an important human factor, and so they believe in the development of educational computer games that individual students' learning needs or difficulties must be considered.

Research predominantly in this phase are the most crucial ones that intend to explore what potential advantage a digital game offers and what roles do digital games have in education. Researchers in this stage are predominantly trying to find answers to the questions, "Can any specific game features be shown to be more effective at supporting learning?"; "Can any game concepts be adapted to suit varying types of subjects and learning styles?"; and "How can an educational computer game be designed effectively to aid student learning?" From a design science perspective, this process can be seen as a design cycle, where digital games as an IT-artefact are constructed, rigorously and thoroughly tested in an experimental situation, and refined further until a satisfactory design is achieved (Hevner, 2007). These aforementioned studies play an important role in providing the foundation of DGBL.

3.2.2. Facilitating learning performance

In the second stage of the development trajectory of DGBL studies, researchers began to bifurcate into three big areas with the main purpose of facilitating learning performance. The first stream focuses on how to make students become engaged, motivated, and have a better experience in an interactive learning environment. The second stream targets how the educational computer game can help students to improve their performance and achievement through different learning strategies. The last stream looks into how DGBL or game concepts can be applied in the classroom to improve classroom teaching.

The researchers in the first stream acknowledge that to be able to gain the greatest potential of DGBL, students need to be engaged, motivated, immersed, and have a better experience in an interactive learning environment (Hamari et al., 2016; Hung et al., 2015; Sung et al., 2017). Thus, much of the research in this area discusses the best approach to enhance students' learning experience in terms of their engagement, learning motivation, and/or immersion. Hamari et al. (2016) and Hung et al. (2015) both argue that challenging games can improve students' learning achievement. Hamari et al. (2016) state that challenging games should be able to keep up with students' growing abilities as a means to keep them maximally engaged in continuous and constant learning. Sung et al.

(2017) report that to increase students' motivations and engagement, experiential learning needs to be integrated into the gaming elements, given that if this aspect is removed from the learning process, then students might not be able to be motivated, to understand the game, or even to interact.

Several studies propose a different game design by integrating learning strategies into gaming scenarios to improve students' performance and learning achievement. Hwang et al. (2013) develop an educational role-playing game and embed the concept maps approach in gaming scenarios and missions. They argue that concept mapping is effective at improving students' learning achievement, especially in natural science courses. Hwang et al. (2015) also focus on improving students' performance and learning achievements. They develop a contextual educational computer game to investigate the effectiveness of an inquiry-based learning strategy on students' learning achievement, learning motivation, degree of satisfaction, and their flow state in a social studies course.

The third stream of research mostly discusses how to incorporate DGBL into in-class activities as instructional tools. For example, Kao et al. (2017) suggest that individual instructors could customize a digital game to achieve their personal instructional goals by administering self-designed learning scaffolds into the game. Zainuddin et al. (2020) investigate the role of e-quizzes on students' learning and engagement. They report that incorporating a game or game concepts into the classroom can be beneficial for educators to retain students' attention and increase their engagement. This view is echoed by Sailer and Sailer (2020), who state that gamified e-quizzes can foster engagement and fun in the classrooms.

4. Discussion

Through main path analysis, we identify two main stages of the DGBL research evolution. The first stage is an exploration of the role of gaming for educational purposes. The early stages are the most crucial ones that intend to explore what potential advantage a digital game offers and what roles do digital game have in education. "Can any specific game features be shown to be more effective at supporting learning?"; "Can any game concepts be adapted to suit varying types of subjects and learning styles?"; and "How can an educational computer game be designed effectively to aid student learning?" are the main questions that the researchers in this stage are trying to answer.

These studies pay great attention to the perceived and actual properties of digital games, which we term "affordance." Strong et al. (2014) define affordances in an organization as *the potential for behaviors [to be] associated with achieving an immediate concrete outcome and arising from the relation between an artifact and a goal-oriented actor or actors*. In this sense, the affordances are the product that arises from the relation of "actors and their goals" (students and/or teacher) and "IT-artefact" (digital game). Acknowledging the affordances that arise from introducing the digital game into education is just the beginning of understanding how this relation implicates the change in an educational organization. Thus, affordances need to be actualized by a goal-oriented actor to achieve an outcome (Strong et al., 2014). Furthermore, Strong et al. (2014) define actualization as *the action taken by actors as they take advantage of one or more perceived affordances through their use of technology to achieve outcomes in support of organizational goals*.

After reviewing the trajectory of knowledge of the DGBL field, we find as a result of the interaction between a goal-oriented actor and digital game in the first phase that researchers then took action to actualize digital game-based learning and achieve an outcome. This outcome is what we find in the second stage of DGBL evolution: facilitate learning performance.

Figure 8 explains that the research focus in the DGBL field is evolving. The arrow indicates the knowledge flow from one work to another work, whereas the thickness of the line denotes its traversal counts.

Over the past decade, Information System researchers have been exploring the affordance actualization lens to understand how information technology is implicated in organizational change processes (Volkoff & Strong, 2017). We believe the change of research topics in the development trajectory of DGBL can be explained better by the theory of affordance actualization. After conducting an extensive literature review, we believe that the affordance actualization used in the field of organization change (Strong et al., 2014) can also be applied to explain the change in DGBL research field. Therefore, we propose a model of Digital Game-based Learning Affordance Actualization (Figure 9).

Figure 8. The transformation of research focus: Actualizing the affordance of digital game-based learning

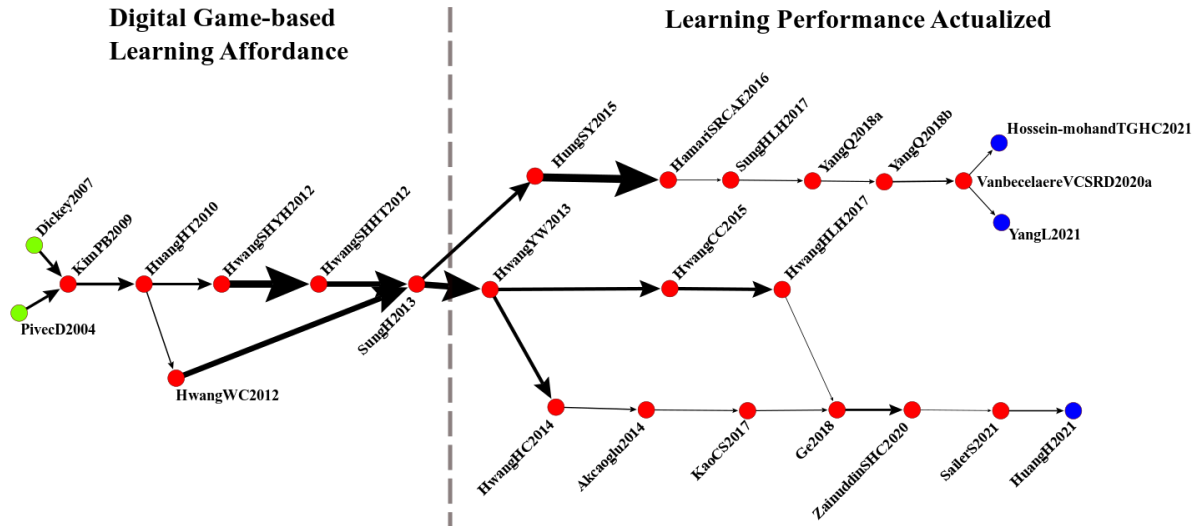
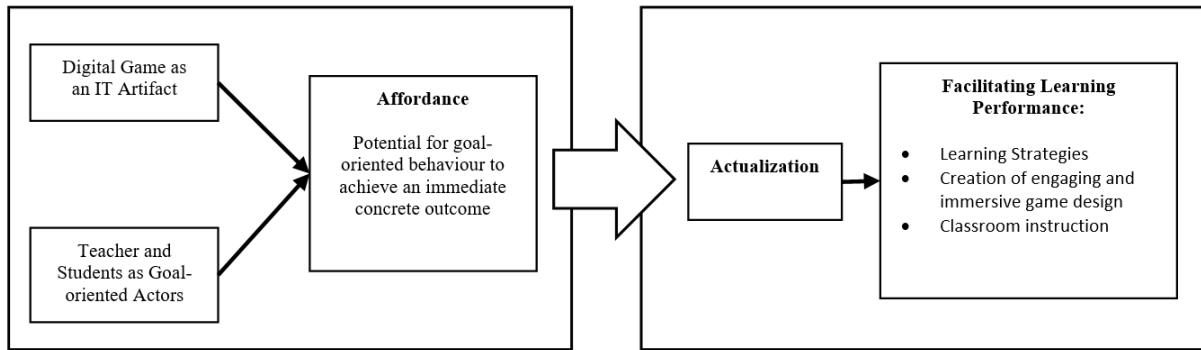


Figure 9. The digital game-based learning affordance actualization model



As we mention above, the first phase of DGBL development trajectories is driven by the desire to enhance the learning environment by presenting new and innovative technology to build digital games for educational purposes. The researchers focus on exploring the potential of digital games and continuously develop them for educational purposes. The second phase of DGBL development trajectories is driven by the effort to actualize DGBL potentials. The researchers look at learning performance through 3 aspects: learning strategies, the creation of engaging and immersive game design, and classroom instruction.

5. Conclusion

The strength and essence of the MPA Approach are that it is able to highlight a series of major developmental events in the field of DGBL research, due to its ability to trace the progress of important works over time. This present study conducts main path analysis to trace the development trajectory of DGBL over a 30-year period. We identify a total of 26 out of 2156 papers related to DGBL as the most significant works.

After analyzing the most significant research that has shaped the development trajectory of DGBL, we present two stages of the evolution of DGBL research focus. In the first stage, most studies focus on developing a game design, feature development, and game concept, while continuously exploring the potential of gaming for educational purposes. Our analysis finds that the evolution of this research focus results from the interaction between a “teacher and students with its goal” and “digital game” in the first phase. Studies then took action to actualize digital game-based learning to achieve an outcome that we note in the second phase of the DGBL evolution. This finding highlights the changing concept of DGBL, which is becoming increasingly complex compared to its early emergence. Thus, understanding the changing concept of DGBL is crucial.

Based on our findings, we propose a model of digital game-based learning affordance actualization. Our proposed model consists of two phases: affordance and actualization. Affordance arises from the relationship between teacher and students as goal-oriented actors and the digital game as an IT artifact. Actualization is the realization of those potentials.

This study contributes to the DGBL literature by offering several distinctions. First, we analyze all the research that has been published related to DGBL since the first time this topic emerged. This fills the gap of previous review paper-related research that only focused on certain aspects of DGBL or were based on a certain period of time. Second, by adopting main path approach, we are able to identify the most significant works in a large and complex citation network of the DGBL literature. Third, knowing the most important juncture in the historical development of DGBL research fields can tell us something about their changing nature in the growth of scientific knowledge. Thus, in this study we explain the evolution of the DGBL literature. An improved understanding of the shift in its evolution may improve the design of learning activities using a game as a better way of learning.

In conclusion, this paper extends prior research by identifying works standing at an important juncture and explains the shift in the DGBL literature through main path analysis. This article presents an extensive literature review with main path lens, which may help new researchers who are considering to enter DGBL research gain insight into what has already been achieved and what should be pursued in future investigations. Understanding the digital game-based learning affordance actualization model will assist instructional designers and educators at developing and creating interactive learning environments by applying a proper learning strategy, creating an engaging and motivating learning environment, and how to construct DGBL into the classroom setting, thus enhancing teaching to facilitate stronger learning performance.

Acknowledgement

The authors are grateful to Professor John S. Liu for his help and advice during the review process. Prof. Liu had made significant intellectual contributions to the study presented in the manuscript.

References

- Acquah, E. O., & Katz, H. T. (2020). Digital game-based L2 learning outcomes for primary through high-school students: A Systematic literature review. *Computers & Education*, 143, 103667. <https://doi.org/10.1016/j.compedu.2019.103667>
- Akcaoglu, M. (2014). Learning problem-solving through making games at the game design and learning summer program. *Educational Technology Research and Development*, 62(5), 583–600. <https://doi.org/10.1007/s11423-014-9347-4>
- Batagelj, V. (2003). Efficient algorithms for citation network analysis. *University of Ljubljana, Institute of Mathematics*, 41(897). 1–29. Preprint Series. <https://doi.org/10.48550/arXiv.cs/0309023>
- Calero-Medina, C., & Noyons, E. C. M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The Case of the absorptive capacity field. *Journal of Informetrics*, 2(4), 272–279. <https://doi.org/10.1016/j.joi.2008.09.005>
- Chen, X., Zou, D., & Xie, H. (2020). Fifty years of British Journal of Educational Technology: A Topic modeling based bibliometric perspective. *British Journal of Educational Technology*, 51(3), 692–708. <https://doi.org/10.1111/bjet.12907>
- Cheng, S. C., Hwang, G. J., & Lai, C. L. (2020). Critical research advancements of flipped learning: A Review of the top 100 highly cited papers. *Interactive Learning Environments*, 0(0), 1–17. <https://doi.org/10.1080/10494820.2020.1765395>
- Chuang, T. C., Liu, J. S., Lu, L. Y. Y., Tseng, F. M., Lee, Y., & Chang, C. T. (2017). The Main paths of eTourism: Trends of managing tourism through Internet. *Asia Pacific Journal of Tourism Research*, 22(2), 213–231. <https://doi.org/10.1080/10941665.2016.1220963>
- Coller, B. D., & Scott, M. J. (2009). Effectiveness of using a video game to teach a course in mechanical engineering. *Computers & Education*, 53(3), 900–912. <https://doi.org/10.1016/j.compedu.2009.05.012>
- Dickey, M. D. (2007). Game design and learning: A Conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development*, 55(3), 253–273. <https://doi.org/10.1007/s11423-006-9004-7>
- Ebner, M., & Holzinger, A. (2007). Successful implementation of user-centered game based learning in higher education: An Example from civil engineering. *Computers and Education*, 49(3), 873–890. <https://doi.org/10.1016/j.compedu.2005.11.026>

- Ge, Z. G. (2018). The Impact of a forfeit-or-prize gamified teaching on e-learners' learning performance. *Computers and Education*, 126(28), 143–152. <https://doi.org/10.1016/j.compedu.2018.07.009>
- Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An Empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54, 170–179. <https://doi.org/10.1016/j.chb.2015.07.045>
- Hevner, A. R. (2007). A Three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), 87–92. <http://aisel.aisnet.org/sjis/vol19/iss2/4>
- Ho, M. H. C., Liu, J. S., & Chang, K. C. T. (2017). To include or not: The Role of review papers in citation-based analysis. *Scientometrics*, 110(1), 65–76. <https://doi.org/10.1007/s11192-016-2158-0>
- Hossein-Mohand, H., Trujillo-Torres, J. M., Gómez-García, M., Hossein-Mohand, H., & Campos-Soto, A. (2021). Analysis of the use and integration of the flipped learning model, project-based learning, and gamification methodologies by secondary school mathematics teachers. *Sustainability*, 13(5), 1–18. <https://doi.org/10.3390/su13052606>
- Huang, B., & Hew, K. F. (2021). Using gamification to design courses: Lessons learned in a three-year design-based study. *Educational Technology & Society*, 24(1), 44–63.
- Huang, W., Huang, W., & Tschopp, J. (2010). Sustaining iterative game playing processes in DGBL: The Relationship between motivational processing and outcome processing. *Computers and Education*, 55(2), 789–797. <https://doi.org/10.1016/j.compedu.2010.03.011>
- Huang, Y. M., & Huang, Y. M. (2015). A Scaffolding strategy to develop handheld sensor-based vocabulary games for improving students' learning motivation and performance. *Educational Technology Research and Development*, 63(5), 691–708. <https://doi.org/10.1007/s11423-015-9382-9>
- Hummon, N. P., & Dereian, P. (1989). Connectivity in a citation network: The Development of DNA theory. *Social Networks*, 11(1), 39–63. [https://doi.org/10.1016/0378-8733\(89\)90017-8](https://doi.org/10.1016/0378-8733(89)90017-8)
- Hung, C. Y., Sun, J. C. Y., & Yu, P. T. (2015). The Benefits of a challenge: student motivation and flow experience in tablet-PC-game-based learning. *Interactive Learning Environments*, 23(2), 172–190. <https://doi.org/10.1080/10494820.2014.997248>
- Hung, H. T., Yang, J. C., & Tsai, Y. C. (2020). Student game design as a literacy practice: A 10-Year review. *Educational Technology & Society*, 23(1), 50–63.
- Hwang, B., Chou, T. C., & Huang, C. H. (2021). Actualizing the affordance of mobile technology for mobile learning: A Main path analysis of mobile learning. *Educational Technology & Society*, 24(4), 67–80.
- Hwang, G. J., Chiu, L. Y., & Chen, C. H. (2015). A Contextual game-based learning approach to improving students' inquiry-based learning performance in social studies courses. *Computers and Education*, 81, 13–25. <https://doi.org/10.1016/j.compedu.2014.09.006>
- Hwang, G. J., Hsu, T. C., Lai, C. L., & Hsueh, C. J. (2017). Interaction of problem-based gaming and learning anxiety in language students' English listening performance and progressive behavioral patterns. *Computers and Education*, 106, 26–42. <https://doi.org/10.1016/j.compedu.2016.11.010>
- Hwang, G. J., Hung, C. M., & Chen, N. S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development*, 62(2), 129–145. <https://doi.org/10.1007/s11423-013-9320-7>
- Hwang, G. J., Sung, H. Y., Hung, C. M., Huang, I., & Tsai, C. C. (2012a). Development of a personalized educational computer game based on students' learning styles. *Educational Technology Research and Development*, 60(4), 623–638. <https://doi.org/10.1007/s11423-012-9241-x>
- Hwang, G. J., Sung, H. Y., Hung, C. M., Yang, L. H., & Huang, I. (2012b). A Knowledge engineering approach to developing educational computer games for improving students' differentiating knowledge. *British Journal of Educational Technology*, 44(2), 183–196. <https://doi.org/10.1111/j.1467-8535.2012.01285.x>
- Hwang, G. J., & Wu, P. H. (2012). Advancements and trends in digital game-based learning research: A Review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology*, 43(1), 6–10. <https://doi.org/10.1111/j.1467-8535.2011.01242.x>
- Hwang, G. J., Wu, P. H., & Chen, C. C. (2012c). An Online game approach for improving students' learning performance in web-based problem-solving activities. *Computers and Education*, 59(4), 1246–1256. <https://doi.org/10.1016/j.compedu.2012.05.009>
- Hwang, G. J., Yang, L. H., & Wang, S. Y. (2013). A Concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers and Education*, 69, 121–130. <https://doi.org/10.1016/j.compedu.2013.07.008>

- Kao, G. Y. M., Chiang, C. H., & Sun, C. T. (2017). Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity. *Computers and Education*, 113, 294–312. <https://doi.org/10.1016/j.compedu.2017.05.022>
- Ke, F. (2008). A Case study of computer gaming for math: Engaged learning from gameplay? *Computers and Education*, 51(4), 1609–1620. <https://doi.org/10.1016/j.compedu.2008.03.003>
- Kim, B., Park, H., & Baek, Y. (2009). Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning. *Computers and Education*, 52(4), 800–810. <https://doi.org/10.1016/j.compedu.2008.12.004>
- Ku, O., Chen, S. Y., Wu, D. H., Lao, A. C. C., & Chan, T. W. (2014). The Effects of game-based learning on mathematical confidence and performance: High ability vs. low ability. *Educational Technology & Society*, 17(3), 65–78.
- Lai, J. W. M., & Bower, M. (2020). Evaluation of technology use in education: Findings from a critical analysis of systematic literature reviews. *Journal of Computer Assisted Learning*, 36(3), 241–259. <https://doi.org/10.1111/jcal.12412>
- Lathabai, H. H., George, S., Prabhakaran, T., & Changat, M. (2018). An Integrated approach to path analysis for weighted citation networks. *Scientometrics*, 117(3), 1871–1904. <https://doi.org/10.1007/S11192-018-2917-1>
- Lin, C. J., Hwang, G. J., Fu, Q. K., & Cao, Y. H. (2020). Facilitating EFL students' English grammar learning performance and behaviors: A Contextual gaming approach. *Computers and Education*, 152, 103876. <https://doi.org/10.1016/j.compedu.2020.103876>
- Liu, J. S., & Lu, L. Y. Y. (2012). An Integrated approach for main path analysis: Development of the Hirsch Index as an Example. *Journal of the American Society for Information Science and Technology*, 63(3), 528–542. <https://doi.org/10.1002/asi.21692>
- Liu, J. S., Lu, L. Y. Y., & Ho, M. H. C. (2019). A Few notes on main path analysis. *Scientometrics*, 119(1), 379–391. <https://doi.org/10.1007/s11192-019-03034-x>
- Martin, F., Dennen, V. P., & Bonk, C. J. (2020). A Synthesis of systematic review research on emerging learning environments and technologies. *Educational Technology Research and Development*, 68(4), 1613–1633. <https://doi.org/10.1007/s11423-020-09812-2>
- Meyer, P. S., Yung, J. W., & Ausubel, J. H. (1999). A Primer on logistic growth and substitution. *Technological Forecasting and Social Change*, 61(3), 247–271. [https://doi.org/10.1016/s0040-1625\(99\)00021-9](https://doi.org/10.1016/s0040-1625(99)00021-9)
- Nooy, W. de, Mrvar, A., & Batagelj, V. (2018). *Explanatory social network analysis with Pajek: Revised and expanded edition for updated software* (3rd ed.). Cambridge University Press. <https://doi.org/10.1017/9781108565691>
- Noroozi, O., Dehghanzadeh, H., & Talaei, E. (2020). A Systematic review on the impacts of game-based learning on argumentation skills. *Entertainment Computing*, 35. <https://doi.org/10.1016/j.entcom.2020.100369>
- Pivec, M., & Dziabenko, O. (2004). Game-based learning in universities and lifelong learning: “UniGame: Social skills and knowledge training” game concept. *Journal of Universal Computer Science*, 10(1), 14–26.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283. <https://doi.org/10.1080/00461520.2015.1122533>
- Quail, N. P. A., & Boyle, J. G. (2019). Virtual patients in health professions education. *Advances in Experimental Medicine and Biology*, 1171, 25–35. https://doi.org/10.1007/978-3-030-24281-7_3
- Sailer, M., & Sailer, M. (2020). Gamification of in-class activities in flipped classroom lectures. *British Journal of Educational Technology*, 52(1), 75–90. <https://doi.org/10.1111/bjet.12948>
- Strong, D. M., Volkoff, O., Johnson, S. A., Pelletier, L. R., Tulu, B., Bar-on, I., Trudel, J., & Garber, L. (2014). A Theory of organization-EHR affordance actualization. *Journal of the Association for Information Systems*, 15(2), 53–85. <https://doi.org/10.17705/1jais.00353>
- Sung, H. Y., & Hwang, G. J. (2013). A Collaborative game-based learning approach to improving students' learning performance in science courses. *Computers and Education*, 63, 43–51. <https://doi.org/10.1016/j.compedu.2012.11.019>
- Sung, H. Y., Hwang, G. J., Lin, C. J., & Hong, T. W. (2017). Experiencing the Analects of Confucius: An Experiential game-based learning approach to promoting students' motivation and conception of learning. *Computers and Education*, 110, 143–153. <https://doi.org/10.1016/j.compedu.2017.03.014>
- Tsai, C. W., & Fan, Y. T. (2013). Research trends in game-based learning research in online learning environments: A Review of studies published in SSCI-indexed journals from 2003 to 2012. *British Journal of Educational Technology*, 44(5), 115–119. <https://doi.org/10.1111/bjet.12031>
- Tüzün, H., Yılmaz-Soylu, M., Karakuş, T., Inal, Y., & Kızılkaya, G. (2009). The Effects of computer games on primary school students' achievement and motivation in geography learning. *Computers and Education*, 52(1), 68–77. <https://doi.org/10.1016/j.compedu.2008.06.008>

Vanbecelaere, S., Van denBerghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2020). The Effects of two digital educational games on cognitive and non-cognitive math and reading outcomes. *Computers and Education*, 143, 103680. <https://doi.org/10.1016/j.compedu.2019.103680>

Volkoff, O., & Strong, D. M. (2017). Affordance theory and how to use it in is research. In *The Routledge Companion to Management Information Systems* (pp. 232–246). <https://doi.org/10.4324/9781315619361>

Yang, J. C., Lin, M. Y. D., & Chen, S. Y. (2018). Effects of anxiety levels on learning performance and gaming performance in digital game-based learning. *Journal of Computer Assisted Learning*, 34(3), 324–334. <https://doi.org/10.1111/jcal.12245>

Yang, J. C., & Quadir, B. (2018). Effects of prior knowledge on learning performance and anxiety in an English learning online role-playing game. *Educational Technology & Society*, 21(3), 174–185.

Za, S., & Spagnoletti, P. (2013). Knowledge creation processes in information systems and management: Lessons from simulation studies. In *Lecture Notes in Information Systems and Organisation* (pp. 191–204). https://doi.org/10.1007/978-3-642-37228-5_19

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The Role of gamified e-quizzes on student learning and engagement: An Interactive gamification solution for a formative assessment system. *Computers and Education*, 145, 103729. <https://doi.org/10.1016/j.compedu.2019.103729>

Appendix

Table A1. Most influential studies in DGBL literature

Label	Author(s)	Title	Subject Area	Participants
PivecD2004	Pivec & Dziabenko (2004)	Game-based learning in universities and lifelong learning: “UniGame: Social Skills and Knowledge Training” game concept	Not specified	Not specified
Dickey2007	Dickey (2007)	Game design and learning: a conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation	Not specified	Not specified
KimPB2009	Kim et al. (2009)	Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning	Not specified	Ninth-grade students
HuangHT2010	Huang et al. (2010)	Sustaining iterative game playing processes in DGBL: The relationship between motivational processing and outcome processing	Economics	Undergraduate students
HwangWC2012	Hwang et al. (2012c)	An online game approach for improving students’ learning performance in web-based problem-solving activities	Natural science course	Fifth and sixth graders of an elementary school
HwangSHHT2012	Hwang et al. (2012a)	Development of a personalized educational computer game based on students’ learning styles	Natural science course	Fifth graders of an elementary school
SungH2013	Sung & Hwang (2013)	A collaborative game-based learning approach to improving students’ learning performance in science courses	Natural science course	Sixth graders of an elementary school
HwangYW2013	Hwang et al. (2013)	A concept map-embedded educational computer game for improving students’ learning performance in natural science courses	Natural science course	Sixth graders of an elementary school
HwangSHYH2013	Hwang et al. (2012b)	A knowledge engineering approach to developing educational computer games for improving students’ differentiating knowledge	Natural science course	Sixth graders of an elementary school

HwangHC2014	Hwang et al. (2014)	Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach	Natural science course	Sixth graders of an elementary school
Akcaoglu2014	Akcaoglu (2014)	Learning problem-solving through making games at the game design and learning summer program	Computer Science	Middle school students
HwangCC2015	Hwang et al. (2015)	A contextual game-based learning approach to improving students' inquiry-based learning performance in social studies courses	Social science course	Sixth graders from an elementary school
HungSY2015	Hung et al. (2015)	The benefits of a challenge: student motivation and flow experience in tablet-PC-game-based learning	Mathematics	Second-grade students
HamariSRCAE 2016	Hamari et al. (2016)	Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning	Quantum physics and engineering dynamics course	High school students and undergraduate mechanical engineering students
HwangHLH2017	Hwang et al. (2017)	Interaction of problem-based gaming and learning anxiety in language students' English listening performance and progressive behavioral patterns	English as a foreign language	Ninth-grade students
SungHLH2017	Sung et al. (2017)	Experiencing the Analects of Confucius: An experiential game-based learning approach to promoting students' motivation and conception of learning	Analects of Confucius in a Chinese course	Fifth graders from an elementary school
KaoCS2017	Kao et al. (2017)	Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity	Physics	Junior high school
YangQ2018a	Yang & Quadir (2018)	Effects of Prior Knowledge on Learning Performance and Anxiety in an English Learning Online Role-Playing Game	English as a foreign language	Sixth graders from an elementary school
YangLC2018	Yang et al. (2018)	Effects of anxiety levels on learning performance and gaming performance in digital game-based learning	English as a foreign language	Fourth graders from an elementary school
Ge2018	Ge (2018)	The impact of a forfeit-or-prize gamified teaching on e-learners' learning performance	English as a foreign language	First-year adult e-learners from an e-learning college
YangQ2018b	Yang & Quadir (2018)	Individual differences in an English learning achievement system: gaming flow experience, gender differences and learning motivation	English as a foreign language	Elementary school students
ZainuddinSHC 2020	Zainuddin et al. (2020)	The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system	Not specified	Junior high students
VanbecelaereV CSRD2020b	Vanbecelaere et al. (2020)	The effects of two digital educational games on cognitive and non-cognitive math and reading outcomes	Math and reading skills	First graders from an elementary school

SailerS2020	Sailer & Sailer (2020)	Gamification of in-class activities in flipped classroom lectures	Not specified	University students
HuangH2021	Huang & Hew (2021)	Using Gamification to Design Courses: Lessons Learned in a Three-year Design-based Study	Undergraduate Introductory Information Management Course	Undergraduate Introductory Information Management students
Hossein-mohandTGHC 2021	Hossein-Mohand et al. (2021)	Analysis of the Use and Integration of the Flipped Learning Model, Project-Based Learning, and Gamification Methodologies by Secondary School Mathematics Teachers	Mathematics	Mathematics teachers

Influences of Growth Mindset, Fixed Mindset, Grit, and Self-determination on Self-efficacy in Game-based Creativity Learning

Yu-chu Yeh^{1,2*}, Yu-Shan Ting³ and Jui-Ling Chiang¹

¹Institute of Teacher Education, National Chengchi University, Taipei City, Taiwan // ²Research Center for Mind, Brain & Learning, National Chengchi University, Taipei City, Taiwan // ³Department of Education, National Chengchi University, Taipei City, Taiwan // ycyeh@nccu.edu.tw // pnm40275@gmail.com // rayechiang@gmail.com

*Corresponding author

(Submitted November 26, 2021; Revised May 31, 2022; Accepted June 13, 2022)

ABSTRACT: Creativity mindset (CM), grit, and self-determination have been defined as critical motivational variables affecting learners' self-efficacy. Therefore, this study pioneers the examination of the relationship between these motivational variables and creativity self-efficacy (CSE) during game-based learning. A Creativity Mindset Inventory (CMI) and a game-based learning intervention were employed. Participants for developing the CMI were 281 3rd to 6th graders, and those for the intervention were 114 3rd and 4th graders. The result revealed that the CMI included four constructs (growth-internal control, growth-external control, fixed-internal control, and fixed-external control). Moreover, the employed intervention enhanced the children's growth CM and CSE. Regression analysis results suggest that self-determination mediates the influence of growth CM and grit on CSE. Additionally, growth CM, especially the growth-internal control CM, is a powerful predictor of self-determination and CSE. In contrast, fixed CM (the overall fixed CM, the fixed-internal control CM, or the fixed-external control CM) does not have any significant influence on self-determination or CSE. Notably, the findings of this study support that growth CM can be enhanced through a well scaffolded educational game. This study contributes to the field of game-based learning by developing a CM inventory, demonstrating a growth CM intervention, and clarifying influential factors to CSE during game-based training. While game-based learning has become popular among elementary school students, the findings of this study provide important insights into the design of game-based learning and creativity training.

Keywords: Creativity, Game-based learning, Growth mindset, Grit, Self-determination

1. Introduction

In well-known theories of creativity (Amabile, 1996; Sternberg & Lubert, 1999), motivation is regarded as a critical element for creative learning. Creativity mindset (CM), grit, and self-determination have been defined as critical motivational variables (e.g., Karwowski, 2014; Hochanadel & Finamore, 2015; Yeh et al., 2020;) affecting learners' self-efficacy. In addition, recent studies have proposed that well-scaffolded digital game-based learning (DGBL) facilitates learning outcomes and motivation effectively (e.g., Bainbridge et al., 2022; Yang & Chen, 2021). We, therefore, tried to examine how these motivational variables stimulate creativity self-efficacy (CSE) during game-based learning.

CSE refers to one's belief in his/her ability to produce creative ideas or solutions and confidence in achieving creative performance (Hass et al., 2016). CM refers to how people perceive their creative ability; it has been divided into the growth and the fixed mindset (Karwowski, 2014). However, identifying more specific types of CM may be required for effective training. Grit, a recently popular concept in psychology, has never been studied in game-based learning; it is defined as the perseverance and passion for long-term goals (Hochanadel & Finamore, 2015; Wang et al., 2018). Self-determination involves the concepts of autonomy, relatedness, and competence (Ryan & Deci, 2000). When self-determination needs are satisfied, personal growth and optimal functioning can be achieved (Millsa et al., 2018).

With the rapid development of technology, creativity has been recognized as a crucial ability (Puccio, 2017). As such, cultivating creativity to adapt to modern society is an imperative educational objective for children. Although many short-term intervention programs have been implemented to enhance children's creativity (e.g., Hoffmann et al. 2021), there is still a relative lack of integration of digital games in creativity training (Yeh et al., 2019; Stolaki & Economides, 2018). Digital game-based learning is effective in stimulating children's problem-solving, critical thinking, and specifically, creativity (Behnamnia et al., 2020; Hooshyar et al., 2019). Such promising game components to foster children include fantasy, curiosity, and challenge (Behnamnia et al., 2020) that target intrinsic motivation among children.

To date, few digital game-based learning interventions have been developed to enhance children's growth CM, even though it has been viewed as a new form of learning with great potential in recent years (Chen et al., 2020; Israel-Fishelson et al., 2021). In the only qualitative study (White & McCoy, 2019) we found, the results showed that students who acquired a growth mindset in creative mathematic game-based learning developed a positive learning attitude and increased their self-efficacy. A well-designed game-based intervention can effectively enhance children's mindful learning, enjoyment, self-determination, and mastery experience while fostering creativity (Yeh et al., 2019, Yeh et al., 2020). These cognitive processes are considered in our digital game-based learning to enhance children's CM. Additionally, no study has examined how children's CM and grit influence their self-determination and, further, affect their CSE during game-based creativity learning. This study attempted to pioneer such research. To achieve our goal, we first developed the Creativity Mindset Inventory (CMI). Then, we employed an intervention of growth CM through game-based learning, by which we investigated the relationships of the concerned variables after the intervention.

2. Theoretical framework

2.1. Creativity mindset constructs

2.1.1. Development of creativity mindset theory

The theory of mindset was originated by examining people's implicit beliefs of intelligence (Dweck, 2007). Based on the malleability and stability of traits, mindsets can be divided into a fixed mindset and a growth mindset (Dweck, 2007; Dweck, 2015). More recently, some researchers have implemented the concept of mindset in creativity studies, which is known as "creativity mindset" (CM) (Karwowski, 2014). CM refers to beliefs or implicit theories about the nature of creativity, and it has been divided into the growth and the fixed creativity mindset (e.g., Hass et al., 2016; Karwowski, 2014; Karwowski et al., 2019; Puente-Díaz & Cavazos-Arroyo, 2019). People with a fixed CM regard creativity as innate and unchangeable. In contrast, people who hold a growth CM see creativity as malleable and able to be developed through learning or practice.

To date, it is still a lack of consensus on whether the growth CM and the fixed CM are two independent constructs or two opposites of the same continuum constructs. O'Connor et al. (2013) considered CM a construct with one end of the continuum constituting the fixed CM and the other the growth CM. On the other hand, some researchers (e.g., Hass et al., 2016; Karwowski, 2014; Karwowski et al., 2019) supported that the fixed and the growth CM are two independent dimensions. Recent studies have shown more evidence supporting the independent-dimension theory (Karwowski et al., 2019; Puente-Díaz & Cavazos-Arroyo, 2017; Zhou et al., 2020). For example, Karwowski (2014) developed a CM inventory that includes two relatively independent yet negatively correlated scales: the growth CM and the fixed CM. Most CM studies have employed such a two-type theory of CM. Moreover, most existing CM inventories have been developed based on adult samples. This study, therefore, sought to identify children's CM and to further understand the relationship between CM and its outcome variables.

2.1.2. An integrated CM theory with learning plasticity and locus of control

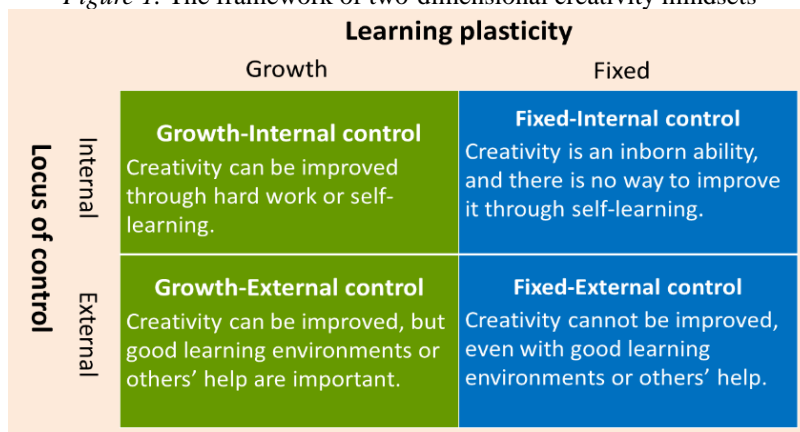
The concept of locus control, derived from the attribution theory, was first proposed by Heider (1958). The attribution theory explains how people interpret the causes of events and how such explanations can be linked with behavior and thinking. Based on a series of attribution studies (Kelley, 1973; Rotter, 1966), Weiner (1985) proposed that causes of success and failure can be divided into three dimensions: locus (internal or external factors), stability (fluctuate or constant), and controllability (controllable or non-controllable). In this study, we tried to integrate the concept of locus of control into our CM theory.

Rotter (1966) defined locus of control as an individual's perception of reinforcement in their life; people interpreted reinforcement in two ways, namely, internal locus of control and external locus of control. Internal control people tend to perceive the outcome as an event depending on their efforts, and they can do anything if they set their mind to it. In contrast, external control people are inclined to believe that the outcome is controlled by external factors, such as chance, fate, and powerful others (Rotter, 1966; Weiner, 1985).

A few researchers have suggested that the concept of mindset overlaps with that of locus of control (Burgoyne et al., 2018; Huillery et al., 2021; Tan et al., 2021). For example, Burgoyne et al. (2018) found a significant relationship between mindset and locus of control. They suggested that growth mindset training could enhance

internal locus of control, challenge-approach motivation, and self-determination. Moreover, internal locus of control was found to be related to creativity performance (Pannells & Claxton, 2008). People with an internal locus of control have a stronger motivation for improvement and try more for getting innovative thoughts and actions than those with an external locus of control (Asgari & Vakili, 2012). These findings advocate that people with growth CM tend to hold an internal locus of control. However, it has been claimed that both internal factors (e.g., knowledge, imagination, attitude, skills) and external factors (e.g., resources, culture, environment, and habitat) are critical to creativity improvement and creativity mindsets (Seelig, 2015; Sternberg & Lubart, 1999; Yeh, 2017). Accordingly, instead of seeing CM as two independent constructs (the growth vs. the fixed mindset), we propose the concept of integrating locus of control into CM. People with different attitudes towards learning plasticity (the growth CM vs. the fixed CM) may simultaneously hold an internal locus of control and an external locus of control. This study, therefore, tried to combine the concepts of learning plasticity and locus of control to develop a more elaborate instrument for measuring CM. Specifically, we propose the following concepts: (1) People who hold a growth-internal control (GI) CM believe that self-learning can improve creativity. (2) People who hold a growth-external control (GE) CM believe that creativity can be enhanced under supportive learning environments or through others' help. (3) People who hold Fixed-Internal control (FI) CM believe that creativity is an inborn ability and that there is no way to improve it through self-learning. (4) People who hold a fixed-external control (FE) CM believe that creativity cannot be improved even under supportive learning environments or with others' help. Notably, the growth CM comprises GI and GE, whereas the fixed CM consists of FI and FE (see Figure 1).

Figure 1. The framework of two-dimensional creativity mindsets



2.2. CM and grit

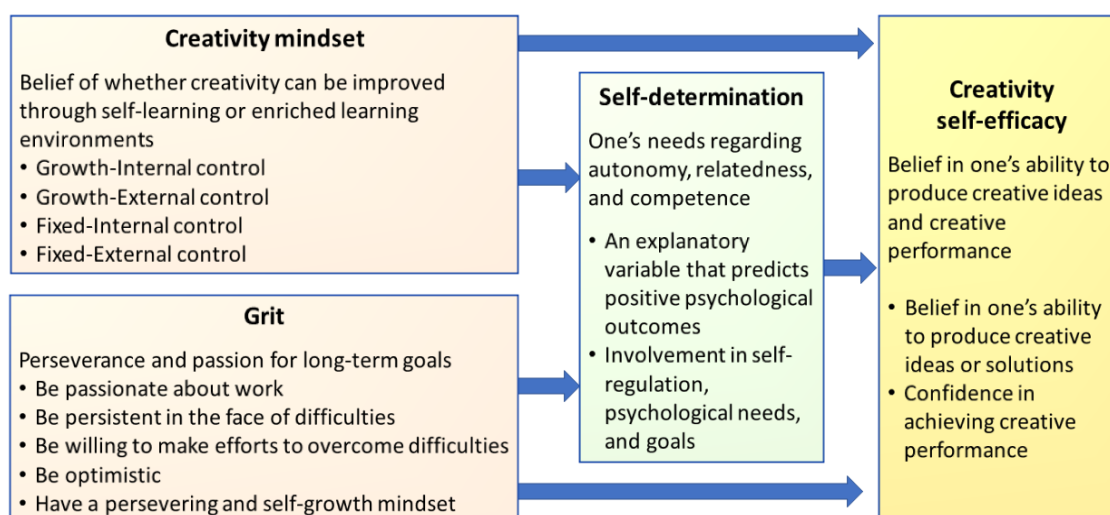
Grit refers to an individual's passion for long-term goals (Duckworth et al., 2007). Duckworth et al. (2007) conceptualized grit as a two-factor structure, namely consistency of interests and perseverance of effort. Empirical findings revealed that grit is related to motivation variables, including future-oriented motivation, self-efficacy, task values, and goal orientations outcomes (Duckworth et al., 2007), especially for grit's perseverance component (Allen et al., 2021; Muenks et al., 2018).

Related studies have shown that individuals with a stronger growth mindset tended to be grittier (Hochanadel & Finamore, 2015). They were more willing to put in efforts to overcome difficulties and had a greater chance of achieving long-term goals (Burgoyne et al., 2018). Similarly, it was found that a growth mindset played an essential role in cultivating a student's trait of grit (Wang et al., 2018). While previous studies have identified the relationship between grit and mindset (Burgoyne et al., 2018; Wang et al., 2018; Zhao et al., 2018), less is known about the association between grit and CM. In studies of personal creativity traits, it has been found that the key concepts of grit (i.e., the passion for long-term goals, consistency of interest, and perseverance of efforts) were central personal traits of creative people (Fisher & Amabile, 2009; Grohman et al., 2017). Creative individuals are persistent and passionate about their work (Fisher & Amabile, 2009); such passion and perseverance successfully predict their creativity (De Clercq et al., 2017). We, therefore, assumed that mindset and grit would interact and then influence the learning of creativity.

2.3. CM, grit, self-determination, and CSE

Creativity self-efficacy (CSE) refers to the belief in one's ability to produce creative ideas or solutions and the confidence in achieving creative performance (Yeh & Lin, 2018; Hass et al., 2016). It has been found that students with a growth mindset usually have stronger motivation to participate and persevere in a task (Zander et al., 2018), whereas children with a fixed mindset have a lower level of self-efficacy (Lee et al., 2022). In the domain of creative studies, it has been suggested that beliefs influence self-perceptions of creativity about the nature of creativity, which involves a person's implicit theory about whether the creative ability is set and unchangeable or can be nurtured. Higher scores on beliefs in the malleability of creativity predicted better scores on a divergent creativity thinking test (O'Connor et al., 2013). These findings suggest that a growth CM may contribute to CSE during game-based creativity learning. Few studies have investigated how grit might be related to children's CSE, especially in the context of game-based learning. In a study investigating children's academic success, Usher et al. (2019) suggested that grit is related to early adolescents' success, particularly when self-efficacy is simultaneously considered. Related studies (Alhadabi & Karpinski, 2019; Muenks et al., 2018) also found that grit correlated positively with students' self-efficacy. We proposed that during game-based creativity learning, grit would help students stay focused and maintain their passion for learning, which would further contribute to CSE. In addition to growth CM and grit, self-determination may influence CSE during game-based creativity learning. On the other hand, fixed CM may have a negative influence on self-determination and CSE. Self-determination has been regarded as a type of intrinsic motivation; it is closely related to self-regulation, psychological needs, and goals (Deci & Ryan, 2008). It has been found that self-determination and self-efficacy are closely related (Martinek & Kipman, 2016).

Figure 2. The theoretical framework of this study



To date, no study has investigated the relationship between CM, grit, self-determination, and CSE during digital game-based learning when interventions of growth CM are employed. Burgoyne et al. (2018) found that measures of mindset, grit, and locus of control loaded onto a common self-determination factor, and the intervention of mindset enhanced learners' growth mindset and self-determination. Participants who received a mindset intervention reported higher scores on growth mindset, internal locus of control, challenge-approach motivation, and self-determination. Similarly, it was found that a growth mindset intervention had a positive influence on the motivation of adolescents (Rhew et al., 2018) and a growth mindset was positively correlated with self-efficacy, task values, and goal orientation (Bai et al., 2021; Dweck, 2007). Additionally, research findings have suggested that growth CM is positively related to creativity performance (Royston & Reiter-Palmon, 2019) and CSE (Puente-Díaz & Cavazos-Arroyo, 2017); in contrast, fixed CM is negatively related to these variables (Karwowski et al., 2019; Puente-Díaz & Cavazos-Arroyo, 2019). Given the aforementioned relationship between growth CM, fixed CM, grit, self-determination, and self-efficacy as well as the negative relationship between fixed CM and growth CM (Hass et al., 2016; Karwowski, 2014; Karwowski et al., 2019; Lee et al., 2022), we assumed that grit and growth CM would enhance CSE directly or indirectly through self-determination, whereas fixed CM would decrease CSE directly or indirectly through self-determination during game-based creativity learning (see Figure 2).

2.4. The present study

To explore the relationship between growth CM, fixed CM, grit, self-determination, and CSE during game-based creativity learning, we developed the Creativity Mindset Inventory and designed a 5-session game-based creativity learning program as the intervention to enhance growth CM and CSE. Empirical findings (Rissanen et al., 2019) have suggested that process focus, mastery orientation, persistence, and individualized student support are core features of growth mindset pedagogy. A recent study (Yeh et al., 2020) has also suggested that the enjoyableness of the game, the encouraging feedback, and the autonomy of gameplay facilitate pupils' motivation and confidence, which further contributes to their improvement of creativity. Therefore, a growth CM and CSE can be built upon mastery and successful experiences. We incorporated these concepts or strategies into our intervention in this study. Notably, since it has been suggested that people with an internal locus of control have a stronger motivation for improvement than those with an external locus of control (Asgari & Vakili, 2012), we assumed that the growth-internal control CM (GI) would be a better predictor of self-determination and CSE than the growth-external control CM (GE). On the other hand, the fixed-internal control CM (FI) would be more detrimental to self-determination and CSE than the fixed-external control CM (FE). The following hypotheses were proposed:

H₁: Growth CM (especially GI) and grit would positively influence self-determination and CSE during game-based creativity learning.

H₂: Fixed CM (especially FI) would negatively influence self-determination and CSE during game-based creativity learning.

H₃: Self-determination would positively influence CSE during game-based learning.

3. Method

3.1. Participants

In developing the CMI, we included 281 3rd to 6th graders (150 boys and 131 girls) from six elementary schools in Taiwan to conduct reliability analysis and confirmatory factor analysis (CFA). Among these pupils, 155 were 3rd and 4th graders (55.2%), and 126 were 5th to 6th graders (44.8%). Written informed consent was obtained from all participants' parents, and each participant was rewarded with a gift valued at 5 USD. In examining path models, 114 3rd and 4th graders (58 boys and 56 girls) from four elementary schools in Taiwan participated in the experimental instruction.

3.2. Instruments

This study employed a game-based creativity learning system and four 6-point Likert type scales (see below) from 1 point to 6 points, representing "strongly disagree" to "strongly agree." Instead of using a 5-point scale, a 6-point Likert scale was employed to avoid the tendency of choosing the middle score of "3." We also designed a reflection questionnaire to understand further the participants' feelings toward the game-based creativity learning program.

3.2.1. Creativity learning system

The learning system of "Digital Game-Based Learning of Creativity-Version A" (DGLC-A), developed for elementary school students (Yeh et al., 2019), was adapted and employed as an instrument to enhance growth CM. The DGLC-A, consisting of nine games, was a story- and game-based learning program. Each game ranged from 10 minutes to 15 minutes. The DGLC-A consisted of the learning of comprehensive creativity strategies and dispositions, such as 3-D creative design, positive thinking and attitude, thinking outside the box, sensitivity in observation, divergent thinking, convergent thinking, lateral thinking, SCAMPER (substitution, combination, adaptation, modification, putting to other uses, elimination, and reversing) and mind mapping (see Figure 3 for example screens). These creativity strategies or dispositions were practiced through 3-D drawing, animations, short stories, open-ended questions, observations, product creation, and problem-solving. We expected that, through mastering creativity skills and positive thinking, the participants would feel self-determined and enhance their growth CM and CSE.

Figure 3. Example screens of the digital game-based learning of creativity-A



3.2.2. Creativity Mindset Inventory (CMI)

The CMI originally included 16 test items with four items in each of the following dimensions: GI, GE, FI, and FE. After reliability and construct validity analysis, one test item in each category was deleted. Twelve test items remained and have good reliability and construct validity (see result session for details). The data was collected in class by the teacher with no time constraints. More details are shown in the results session.

3.2.3. The Grit Scale

An adapted Grit Scale was employed to measure the participants' trait of grit. The original Grit Scale, with 12 items, was developed by Duckworth et al. (2007). With permission, the Grit Scale was translated, adapted, and validated by reliability and factor analysis based on 338 3rd to 6th graders. Four items were deleted after exploratory factor analysis and reliability analysis. The adapted Grit Scale included two factors: perseverance of effort (4 items) and consistency of interest (4 items). The test items included statements such as "I finish whatever I begin" and "New ideas and projects won't distract me from previous ones." The Cronbach's α coefficients for the whole inventory and two factors (perseverance of effort and consistency of interest) were .906, .872, and .813, respectively (Yeh, 2020).

3.2.4. Inventory of self-determination in digital games

The Inventory of Self-Determination in Digital Games (ISD-DG) (Yeh et al., 2019) was employed to measure the participants' level of self-determination during the game-based creativity learning. The ISD-DG, with 13 items, consists of two factors: autonomy and self-regulation (7 items) and competence (6 items). The test items included statements such as "I had many chances to make free choices" and "I could achieve the scores or goals that I set." The Cronbach's α coefficients for the two factors and the total score of the IDS-DG were .887, .881, and .933, respectively.

3.2.5. Inventory of self-efficacy in creativity digital games

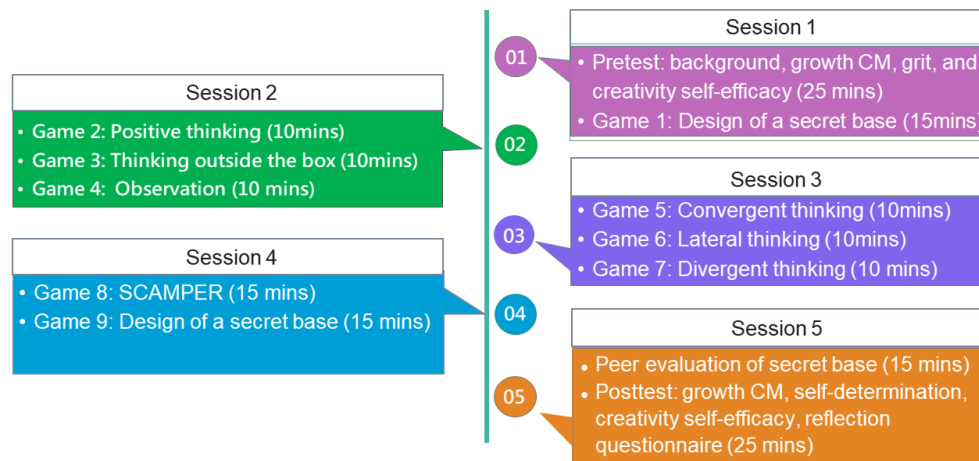
The Inventory of Self-Efficacy in Creativity Digital Games (IS-CDG) (Yeh & Lin, 2018) was employed to measure the participants' level of self-efficacy after game-based creativity learning. The IS-CDG contains nine items, including two factors: the ability to generate creative ideas (6 items) and achievement of creative performance (3 items). The test items included statements such as "I believe that I can come up with many creative ideas" and "I am more creative than most of my classmates." The Cronbach's α coefficients for the two factors and the total score of the ISE-DG were .908, .844, and .927, respectively.

3.3. Experimental design and procedures

All participants completed the experiment in the computer laboratory at their school during their flexible learning time or computer class. The participants were asked to complete nine games in the DGLC-A in 5 class sessions within a week; each session was 40 minutes. Before starting the session, all participants took the pretest, including background information, growth CM, grit, and CSE. After completing the DGLC-A, students were asked to conduct peer evaluation, then completed the posttest, which included growth CM, self-determination, CSE, and the reflection questionnaire. Participants of the same class completed each session as a group (see Figure 4).

Aside from embedding strategies to boost students' creative ability and dispositions, the features of DGLC-A also incorporated other instructional strategies, including scaffolding to challenge their creativity skills, offering chances for self-determination (free choice of game order), providing constructive feedback for answers, utilizing verbal encouragement for performance, and providing peer evaluation for creative design. Peer evaluations were employed to rate the popularity and creativity of the designed products in game 1 and game 9, during which observational learning was expected. These teaching strategies were employed to potentially enhance the participants' growth CM, which is in line with the suggestions that mindset can be enhanced through process focus, mastery orientation, persistence, and individualized student support (Rissanen et al., 2019). Specific experimental procedures are illustrated in Figure 4.

Figure 4. Procedures and interventions for the experiment



4. Results

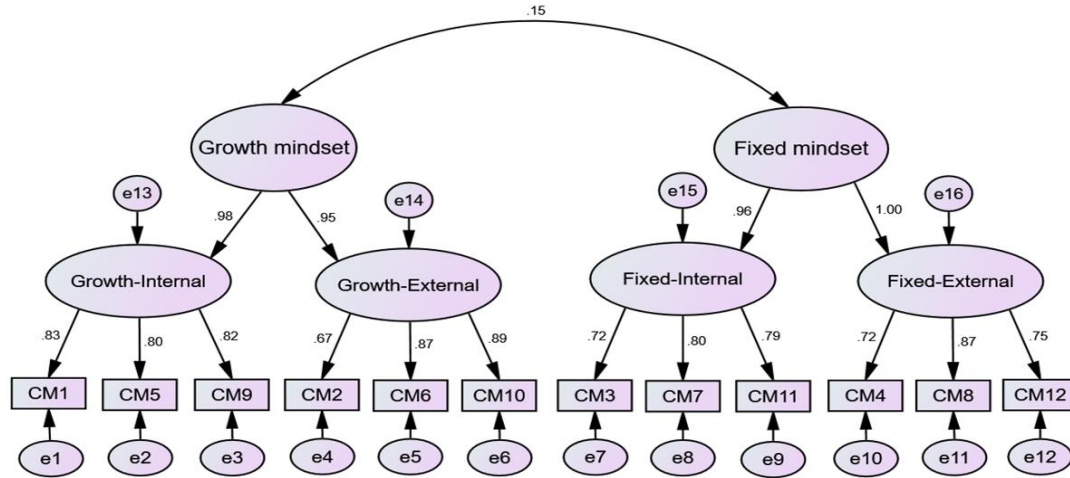
4.1. The development of the CMI

The CMI includes the growth CM (GE and GI) and the fixed CM (FI and FE). The exploratory factor analysis (EFA) was first conducted to examine the construct validity of the CMI using a random split-half of the sample ($N = 135$, 62 boys and 73 girls). Then, the confirmatory factor analysis was employed to validate the CMI using the second split-half sample ($N = 146$, 88 boys and 58 girls). Principal Component Analysis and direct varimax were employed in factor extraction and rotation when conducting EFA (see Table Appendix 1). With factor loadings ranging from .409 to .909, 71.22% of the total variance was explained by GI and GE, and 86.03% of the total variance was explained by FI and FE. Regarding internal-consistency reliability, the Cronbach's α coefficients for growth CM, GI, and GE were .911, .859, and .850, respectively. The Cronbach's α for the fixed CM, FI, and FE were .952, .877, and .924, respectively. Moreover, the item-total correlation coefficients ranged from .622 to .907.

A second-order CFA model (see Figure 5) was examined based on variance-covariance matrices and maximum likelihood estimation through Amos. The following criteria were employed to examine the model fit: a non-significant chi-square degree of freedom ratio (χ^2/df), the comparative fit index (CFI) higher than .90, the root mean square error of approximation (RMSEA) lower than .10, and the standard root mean squared residual (SRMR) less than .08 (Iacobucci, 2010; Kenny et al., 2015). Our CFA results were as follows: $\chi^2 (N = 146, df = 51) = 107.832, p < .001$, the SRMR = .070, the RMSEA = .088, and the CFI = .947. Due to the ratio of χ^2 is sensitive to the sample size, $\chi^2/df \leq 3$ is acceptable (Iacobucci, 2010). The composite reliability (ρ_c) for GI, GE,

FI, and FE were .86, .86, .82, and .83, respectively. The average variance extracted (ρ_v) values for the four factors were .67, .67, .60, and .62, respectively. These results support that the CMI has good reliability and construct validity; moreover, M is composed of growth and fixed CM, with two sub-types of CM (internal-control and external control) under each construct.

Figure 5. Confirmatory factor analysis results of the inventory of creativity mindset



Lastly, using all the samples ($N = 281$) to conduct Pearson correlation analysis, we found that the total score of the growth CM and the fixed CM were slightly correlated ($r = .213, p < .001$). GI and GE were moderately correlated ($r = .433, p < .001$). While GE was moderately related to FI or FE, GI was not related to any type of fixed mindset. On the other hand, FI and FE were highly correlated ($r = .841, p < .001$) (see Table 1).

Table 1. The correlations among the growth CM, the fixed CM, and the four sub-types of CM

Variable	FI	FE	Fixed CM	GI	GE	Growth CM
FI	1					
FE	.841***	1				
Fixed CM	.958***	.961***	1			
GI	-.046	-.054	-.052	1		
GE	.400***	.377***	.405***	.433***	1	
Growth CM	.214***	.195**	.213***	.841***	.852***	1

Note. ** $p < .01$; *** $p < .001$.

4.2. Preliminary analysis of intervention

Since the relationships of the concerned variables were investigated through the game-based learning intervention we developed, it was necessary to examine whether the vehicle was effective. Therefore, we conducted a repeated measure analysis of variance to separately examine whether the participants enhanced their growth CM (GI and GE) and CSE after the game-based creativity learning. The results showed that the participants' growth CM had leveled up, $F(1, 113) = 7.463, p = .007, \eta^2_p = .062$, and $F(1, 113) = 8.614, p = .004, \eta^2_p = .071$ for GI and GE, respectively. In addition, the results showed that the participants' overall CSE had been enhanced, $F(1, 113) = 4.860, p = .030, \eta^2_p = .041$ (see Figure 6 for Ms and SEs). These findings suggest that the game-based learning intervention was effective.

4.3. Relationship of CM, grit, self-determination, and CSE

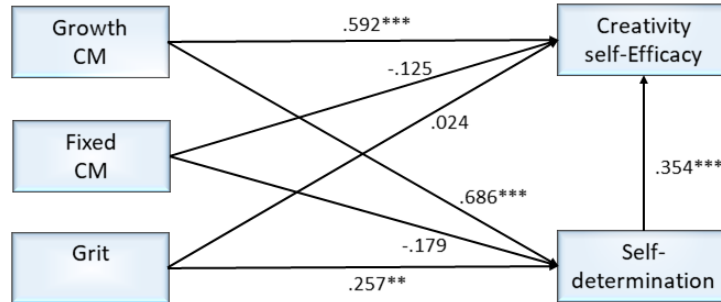
To investigate the relationships between grit, CM, self-determination, and CSE after game-based learning, we conducted stepwise multiple regression analyses. The factor loadings, Cronbach's alpha, and composite reliability (CR) were assessed for the internal consistency; the convergent validity of the scales based on the average variance extracted (AVE) was also measured (see Table 2). A mean-centered approach was employed for each construct prior to the analysis to support the use of all the information (Marsh et al., 2007).

Results of stepwise regression analyses revealed that when using growth CM, fixed CM, and grit to predict self-determination, only GI and grit could significantly predict self-determination, $F(1, 111) = 48.165, p < .001$; the

variance explained was 46.5 %. When using growth CM, fixed CM, grit, and self-determination to predict CSE, only growth CM and self-determination could significantly predict CSE, $F(2, 111) = 91.064, p < .001$; the variance explained was 61.5 % (see Table 3). Figure 6 visualized the results from the regressions analyses.

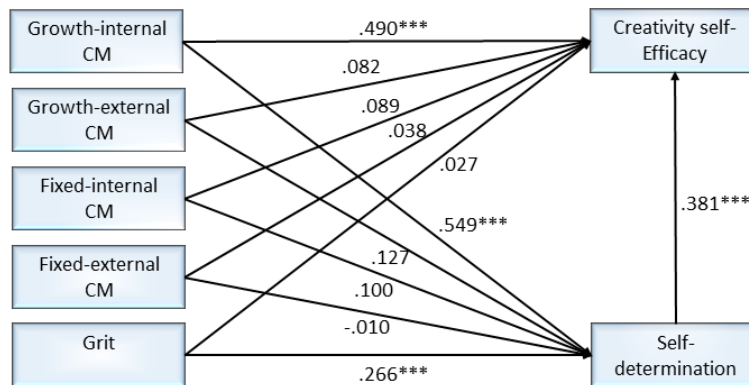
When using the two constructs of growth CM, the two constructs of fixed CM, and grit to predict self-determination, only GI and grit could significantly predict self-determination, $F(1, 111) = 48.307, p < .001$; the variance explained was 46.5 %. When using the two constructs of growth CM, the two constructs of fixed CM, grit, and self-determination to predict CSE, only GI and self-determination could significantly predict CSE, $F(2, 111) = 91.398, p < .001$; the variance explained was 61.5 %. (see Table 4). Figure 7 visualized the results from the regressions analysis.

Figure 6. The standardized regression coefficients of the overall growth CM, fixed CM, grit, and self-determination on CSE



Note. ** $p < .01$. *** $p < .001$.

Figure 7. The standardized regression coefficients of the two constructs of growth CM, the two constructs of fixed CM, grit, and self-determination on CSE



Note. ** $p < .01$. *** $p < .001$.

Table 2. The factor loadings, Cronbach's alpha, CR, and AVE values of the model

Construct	Factor loadings	Cronbach's α	CR	AVE
Growth CM	.699 to .833	.873	.904	.613
Growth-external control (GE)	.784 to .835	.759	.861	.675
Growth-internal control (GI)	.846 to .880	.827	.896	.743
Fixed CM	.854 to .935	.958	.966	.828
Fixed-external control (FE)	.924 to .947	.925	.952	.869
Fixed-internal control (FI)	.898 to .934	.904	.940	.839
Grit	.458 to .833	.850	.886	.501
Perseverance	.549 to .854	.727	.830	.555
Consistency of interest	.572 to .789	.769	.854	.600
Self-determination	.716 to .844	.955	.960	.651
Autonomy & self-regulation	.717 to .869	.915	.932	.665
Competency	.761 to .868	.915	.934	.703
Creativity self-efficacy	.808 to .907	.956	.962	.739
Ability	.838 to .905	.947	.958	.791
Achievement	.926 to .943	.926	.953	.872

Table 3. Result of multiple regression analyses with the overall growth CM and fixed CM

Model	IVs	β	t	p	VIF	R	R^2	F	R^2 Change	F change
Self-determination as the dependent variable										
1	Growth CM	.635	8.702***	.000	1.000	.635	.403	75.732***	.403	75.732***
2	Growth CM	.550	7.481***	.000	1.119	.682	.465	48.165***	.061	12.692***
	Grit	.262	3.563***	.001	1.119					
CSE as the dependent variable										
1	Growth CM	.731	11.343***	.000	1.000	.731	.535	128.658***	.535	128.658***
2	Growth CM	.489	6.467	.000	1.676	.788	.621	91.064***	.087	25.420***
	Self-determination	.381	5.042	.000	1.676					

Note. *** $p < .001$.

Table 4. Result of multiple regression analyses with factors of growth CM and fixed CM

Model	IVs	β	t	p	VIF	R	R^2	F	R^2 Change	F change
Self-determination as the dependent variable										
1	GI	.686	6.366***	.000	1.000	.634	.402	75.205***	.402	75.205***
2	GI	.549	7.496***	.000	1.113	.682	.465	48.307***	.064	13.210***
	Grit	.266	3.635***	.000	1.113					
CSE as the dependent variable										
1	GI	.732	11.355***	.000	1.000	.732	.530	128.940***	.535	128.940***
2	GI	.490	6.494***	.000	1.671	.789	.615	91.398***	.087	25.569***
	Self-determination	.381	5.057***	.000	1.671					

Note. *** $p < .001$.

5. Discussion

5.1. Development of the CMI

Mindset is typically divided into a growth mindset and a fixed mindset (Dweck, 2007). Some researchers assume that the growth mindset and the fixed mindset are independent factors (e.g., Karwowski, 2014), whereas some postulate that people may endorse both mindsets depending on circumstances (Hass et al., 2016). In this study, we propose a 2-dimensional CM theory (learning plasticity and locus of control) in which four types of mindsets are identified: Growth-Internal control (GI), Growth-External control (GE), Fixed-Internal control (FI), and Fixed-External control (FE). The results of this study suggest that the CMI has good reliability and construct validity. Additionally, the results of second-order CFA support our two-dimensional constructs of CM. The results support the claim that mindset overlaps with that of locus of control (Huillery et al., 2021; Tan et al., 2021), growth mindset and internal locus of control are related (Price et al., 2018), and both internal factors and external factors are critical to creativity improvement and creativity mindsets (Seelig, 2015; Yeh, 2017).

Correlation analyses suggest that the overall growth CM and fixed CM have a low positive correlation. However, while GE has a moderate positive relation with FI and FE, GI did not have such a positive relation. These results manifest the importance of our attempt to identify specific types of CM beyond overall growth and fixed CM. The findings suggest that children may simultaneously hold the growth CM and the fixed CM. Although these two concepts are relatively independent (Puente-Díaz & Cavazos-Arroyo, 2019), they are not necessarily the opposite. The results support Karwowski's (2014) argument that people can hold both an entity and an incremental view of creativity; they can be convinced that great creators are enabled by an inborn power and agree that personal effort can increase their creative potential.

Moreover, this study found that the belief that CM can be improved in an enriched environment or through others' help (i.e., Growth-External control CM), in a way, overlaps with fixed CM. This result does not support the findings of Hass et al. (2016), in which fixed and growth mindsets were negatively related in a college student sample. Our participants were 3rd and 4th-grade students. Previous findings that background factors (e.g., age, life experience) affect people's locus of control development (Cummins & Nistico, 2002; Pannells & Claxton, 2008) may explain the difference. The positive relationship between growth-external CM and fixed CM may imply that children perceive external resources as restrictions they cannot control. How to transform such a mindset of limitation into resources has become vital. Altogether, the findings of this study suggest that, although the four types of CM can be explained by two factors (growth CM and fixed CM), the four-factor structure can better describe children's CM and children's beliefs about growth fixed CM co-exist.

5.2. Relationship of growth CM, fixed CM, grit, self-determination, and CSE

Before testing our hypothesis regarding the relationship between growth CM, fixed CM, grit, self-determination, and CSE during game-based learning, we examined the effects of the game-based intervention. This process ensures a valid intervention and provides a reliable basis for our further investigation of the relationship among the concerned variables. The results suggest that our incorporated strategies (i.e., task design, scaffolding, self-determination opportunities, constructive and immediate feedback, verbal encouragement, and peer evaluation) in the DGLC-A successfully boost the children's growth CM and CSE. These features effectively enhanced pupils' growth CM and CSE during game-based learning. The results also align with past findings that game-embedded animations effectively promote conceptual understanding (Bainbridge et al., 2022), and adequately integrating learning strategies into digital games can effectively improve students' learning achievement (Yang & Chen, 2021).

In this study, we proposed three hypotheses to examine the relationship between growth CM, fixed CM, grit, self-determination, and CSE. The findings of multiple regression analyses suggest that growth CM (especially GI) and grit positively influence self-determination and CSE during game-based learning; moreover, self-determination positively influences CSE during game-based learning. These results support our hypotheses 1 and 3. However, our hypothesis 2 is not supported. We found that fixed CM could not predict self-determination or CSE, which is consistent with Karwowski's (2014) finding. However, the finding is contradictory to Lee et al.'s (2022) finding that children's fixed mindset negatively influences their self-efficacy. Specifically, the findings of this study suggest that growth CM (especially the GI) is a strong predictor of self-determination and CSE, whereas the overall fixed CM or the two constructs of fixed CM cannot predict self-determination or CSE during game-based learning. These results also suggest that enhancing growth-internal CM is critical to pupils' development of CSE.

To date, no study has examined the relationship between growth CM, grit, self-determination, and CSE during game-based learning. The relationship between grit, growth CM (especially GI), and CSE found in this study support previous findings that growth mindset and self-efficacy are related (Karwowski & Kaufman, 2017; Price et al., 2018), and grit is positively related to self-growth mindset (Hochanadel & Finamore, 2015). Our findings also support that a growth mindset and grit are interconnected dispositions (Keesey et al., 2018), grit correlated positively with students' self-efficacy (Alhadabi & Karpinski, 2019; Muenks et al., 2018), and a growth mindset, grit, and SD are closely associated (Burgoyne et al., 2018). However, we found that growth CM (especially GI) was a more important predictor of self-determination and CSE than grit after the game-based learning.

Notably, the findings in the regression models suggest that self-determination serves as a mediator of growth CM and creativity self-efficacy during game-based creativity learning. Two major indicators of self-determination are autonomy and competence (Ryan & Deci, 2000). When people believe that creativity can be improved, they may be more confident and willing to take challenges to pursue autonomy and obtain competencies during game-based learning, by which their CSE is enhanced. These results are in line with past findings that self-determination and self-efficacy are positively related (Develos-Sacdalan & Bozkus, 2018) and that grit is important in learning outcomes (Duckworth, & Quinn, 2009; Muenks et al., 2018). The results also support that self-determination (Millsa et al., 2018) is critical to the effectiveness of game-based learning. When self-determination needs are satisfied, personal growth and optimal functioning can be achieved (Millsa et al., 2018).

6. Conclusions

As creativity is crucial to future success and growth CM is critical to creative learning, there is a need to develop an enjoyable growth CM intervention to help children build up their CSE. Meanwhile, understanding influential factors in such interventional learning is essential. The existing construct of CM (the growth vs. the fixed CM) may not be specific enough to identify children's CM and, accordingly, provide effective interventions. Therefore, we proposed four types of CM (GE, GI, FI, and FE) under the growth and the fixed CM construct, by which we developed the Creativity Mindset Inventory (CMI) based on a 2-dimensional CM (learning plasticity and locus of control) theory and developed a game-based learning intervention. How growth CM, fixed CM, grit, and self-determination may influence CSE was examined. The results suggest that the CMI is a valid instrument for measuring children's CM, and it can help distinguish children's specific beliefs toward CM.

In addition, this study contributes to game-based learning by clarifying the relationships among different types of growth CM and fixed CM, grit, self-determination, and CSE during game-based creativity learning. The results suggest that self-determination is a vital mediator between the concerned variables, which provides evidence for

learning processes. This study also sheds light on how growth CM (especially GI) can be improved to enhance CSE in game-based creativity learning through embedded concrete instructional strategies. As game-based learning has become popular among elementary school students during the COVID-19 pandemic era, the findings of this study provide important insights into the design of game-based learning and creativity training.

7. Limitations and implications

Because the perceptions of self-determination during game playing cannot be measured before the intervention, self-determination's learning effect was not examined. Further studies can extend the intervention and measure self-determination at different time points, by which the dynamic influence of self-determination can be added to path models. Moreover, owing to the difficulty of convincing elementary schools to allocate more experimental time, only five sessions (40 mins each) of training were employed. A longer experimental duration may better enhance the growth CM and CSE. Nevertheless, the positive learning effect of this short intervention was confirmed through the repeated measure analysis of variance and the responses from the reflection questionnaire. Further studies can also include a control group to double-check the learning effect if enough participants are recruited.

In this study, we identified four types of CM and accordingly developed the CMI, which serves as an effective instrument for measuring CM. Moreover, this study found that pupils' beliefs of growth-internal ($M = 4.44$, $SD = 1.09$) and growth-external CM ($M = 3.99$, $SD = 1.22$) were much stronger than that of fixed-internal ($M = 3.12$, $SD = 1.23$) and fixed-external CM ($M = 2.81$, $SD = 1.41$), suggesting that children are optimistic toward their creative development and that there is great learning plasticity if an enriched environment can be provided. Therefore, instructors or researchers can use the CMI to obtain specific information about learners' beliefs of different types of growth or fixed CM, by which effective training or instruction can be designed to maximize learning effects.

Furthermore, growth CM promotes a positive attitude and willingness to try new ideas and new things. The strong influence of growth CM (especially GI) on CSE through self-determination suggests that when children believe that creativity can be improved through self-learning in a well-facilitated environment, they may be more autonomous and competent during game-based learning. As a result, they may become substantially more creative. The positive intervention results of this study suggest that developing effective interventions to enhance growth CM through game-based learning is an efficient and enjoyable way to achieve this goal. Researchers in education and game designers can cooperate in developing more game-based learning programs to enhance the growth CM, especially GI.

Self-efficacy is a vital precursor to successful performance (e.g., Schunk & DiBenedetto, 2016), and a growth CM is positively related to CSE (Karwowski & Kaufman, 2017). This study suggests that incorporating strategies such as scaffolding for challenging their creativity skills, chances for self-determination, constructive and immediate feedback, verbal encouragement for performance, and peer evaluation for creative design can enhance growth CM and CSE. These strategies can also be implemented in classroom teaching. Moreover, we enhanced the children's growth CM and CSE mainly through practicing creative strategies and dispositions in this study. Future studies can incorporate more strategies for enhancing growth CM in game-based learning.

Finally, different from past related studies, we identified four types of CM under two constructs (growth CM and fixed CM). We found that the growth-internal CM has stronger correlations with the other concerned variables than the growth-external control. People who hold a growth-internal control CM believe that self-learning can improve creativity; such a belief is more important than ever during the COVID-19 pandemic. This worldwide pandemic has revealed the importance of self-learning through digital vehicles. Our development of the CM instrument and the digital game-based intervention, which can be completed through self-learning, provides unique contributions and implications in this critical era.

Acknowledgment

This work was supported by the Ministry of Science and Technology in Taiwan under Grant NSC MOST 107-2410-H-004 -079 -SS2. We thank Han-Lin Chang and Yu-Jung Lin for helping revise the game-based learning system.

References

- Alhadabi, A., & Karpinski, A. C. (2019). Grit, self-efficacy, achievement orientation goals, and academic performance in university students. *International Journal of Adolescence and Youth*, 25(1), 519-535. <https://doi.org/10.1080/02673843.2019.1679202>
- Allen, R. E., Kannangara, C., & Carson, J. (2021). True grit: How important is the concept of grit for education? a narrative literature review. *International Journal of Educational Psychology*, 10(1), 73-87. <https://doi.org/10.17583/ijep.2021.4578>
- Amabile, T. M. (1996). *Creativity in context*. Westview Press.
- Asgari, M. H., & Vakili, M. (2012). The Relationship between locus of control, creativity, and performance of the educational department employees in the west of Mazandaran. *International Research Journal of Applied and Basic Sciences*, 3(S), 2556-2561.
- Bai, B., Wang, J., & Nie, Y. (2021). Self-efficacy, task values and growth mindset: what has the most predictive power for primary school students' self-regulated learning in English writing and writing competence in an Asian Confucian cultural context? *Cambridge Journal of Education*, 51(1), 65-84. <https://doi.org/10.1080/0305764X.2020.1778639>
- Bainbridge, K., Shute, V., Rahimi, S., Liu, Z., Slater, S., Baker, R. S., & D'Mello, S. K. (2022). Does embedding learning supports enhance transfer during game-based learning? *Learning and Instruction*, 77, 101547. <https://doi.org/10.1016/j.learninstruc.2021.101547>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215.
- Behnamnia, N., Kamsin, A., Ismail, M. A. B., & Hayati, A. (2020). The Effective components of creativity in digital game-based learning among young children: A Case study. *Children and Youth Services Review*, 116, Article 105227. <https://doi.org/10.1016/j.chidyouth.2020.105227>
- Burgoyne, A., Hambrick, D., Moser, J., & Burt, A. (2018). Analysis of a mindset intervention. *Journal of Research in Personality*, 77, 21-30. <https://doi.org/10.1016/j.jrp.2018.09.004>
- Cummins, R. A., & Nistico, H. (2002). Maintaining life satisfaction: The Role of positive cognitive bias. *Journal of Happiness studies*, 3(1), 37-69. <https://doi.org/10.1023/A:1015678915305>
- De Clercq, D., Mohammad Rahman, Z., & Belausteguigoitia, I. (2017). Task conflict and employee creativity: The Critical roles of learning orientation and goal congruence. *Human Resource Management*, 56(1), 93-109. <https://doi.org/10.1002/hrm.21761>
- Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A Macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie canadienne*, 49(3), 182-185. <https://doi.org/10.1037/a0012801>
- Develos-Sacalan, K., & Bozkus, K. (2018). The Mediator role of resilience between self-determination and self-efficacy. *Education Science and Psychology*, 4(50), 49-60.
- Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT-S). *Journal of Personality Assessment*, 91(2), 166-174. <https://doi.org/10.1080/00223890802634290>
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92, 1087-1101. <https://doi.org/10.1037/0022-3514.92.6.1087>
- Dweck, C. S. (2007). *Mindset: The New psychology of success*. Random House.
- Dweck, C. S. (2015). Carol Dweck revisits the growth mindset. *Education Week*, 35(5), 20-24.
- Fisher, C. M., & Amabile, T. (2009). Creativity, organization, and improvisation. In T. Rickards, M. A. Runco, & S. Moger (Eds.), *The Routledge companion to creativity* (pp. 11-24). Routledge. <https://doi.org/10.4324/9780203888841.pt2>
- Grohman, M. G., Ivcevic, Z., Silvia, P., & Kaufman, S. B. (2017). The Role of passion and persistence in creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 11(4), 376-385. <https://doi.org/10.1037/aca0000121>
- Hass, R. W., Katz-Buonincontro, J., & Reiter-Palmon, R. (2016). Disentangling creative mindsets from creative self-efficacy and creative identity: Do people hold fixed and growth theories of creativity? *Psychology of Aesthetics, Creativity, and the Arts*, 10(4), 436-446. <https://doi.org/10.1037/aca0000081>
- Heider, F. (1958). *The Psychology of interpersonal relations*. John Wiley & Sons Inc. <https://doi.org/10.1037/10628-000>
- Hochanadel, A., & Finamore, D. (2015). Fixed and growth mindset in education and how grit helps students persist in the face of adversity. *Journal of International Education Research*, 11(1), 47-50. <https://doi.org/10.19030/jier.v11i1.9099>
- Hoffmann, J. D., Ivcevic, Z., & Maliakkal, N. (2021). Emotions, creativity, and the arts: Evaluating a course for children. *Empirical Studies of the Arts*, 39(2), 123-148. <https://doi.org/10.1177/0276237420907864>

- Hooshyar, D., Kori, K., Pedaste, M., & Bardone, E. (2019). The Potential of open learner models to promote active thinking by enhancing self-regulated learning in online higher education learning environments. *British Journal of Educational Technology*, 50(5), 2365-2386. <https://doi.org/10.1111/bjet.12826>
- Huillery, E., Bouguen, A., Charpentier, A., Algan, Y., & Chevallier, C. (2021). The role of mindset in education: A large-scale field experiment in disadvantaged schools. <https://doi.org/10.31235/osf.io/zs9aq>
- Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of consumer psychology*, 20(1), 90-98. <https://doi.org/10.1016/j.jcps.2009.09.003>
- Karwowski, M. (2014). Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1), 62-70. <https://doi.org/10.1037/a0034898>
- Karwowski, M., & Kaufman, J. C. (Eds.). (2017). *The Creative self: Effect of beliefs, self-efficacy, mindset, and identity*. Academic Press.
- Karwowski, M., Lebeda, I., & Beghetto, R. A. (2019). Creative self-beliefs. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (2nd ed., pp. 396-417). Cambridge University Press.
- Keesey, S., Schaefer, A., Loy, M., & Allen, C. J. (2018). Developing growth mindset and GRIT in preservice teachers. *Kentucky Teacher Education Journal: The Journal of the Teacher Education Division of the Kentucky Council for Exceptional Children*, 5(1), Article 3. <https://digitalcommons.murraystate.edu/ktej/vol5/iss1/3>
- Kelley, H. H. (1973). The Processes of causal attribution. *American Psychologist*, 28(2), 107-128. <https://doi.org/10.1037/h0034225>
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486-507. <https://doi.org/10.1177/0049124114543236>
- Lee, H. J., Lee, J., Song, J., Kim, S., & Bong, M. (2022). Promoting children's math motivation by changing parents' gender stereotypes and expectations for math. *Journal of Educational Psychology*, 114(7), 1567-1588. <https://doi.org/10.1037/edu0000743>
- Marsh, H. W., Wen, Z., Hau, K. T., Little, T. D., Bovaird, J. A., & Widaman, K. F. (2007). Unconstrained structural equation models of latent interactions: Contrasting residual- and mean-centered approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 570-580. <https://doi.org/10.1080/10705510701303921>
- Millsa, D. J., Milyavskaya, M., Mettler, J., & Heath, N. L. (2018). Exploring the pull and push underlying problem video game use: A Self-determination theory approach. *Personality and Individual Differences*, 135, 176-181. <https://doi.org/10.1016/j.paid.2018.07.007>
- Muenks, K., Yang, J. S., & Wigfield, A. (2018). Associations between grit, motivation, and achievement in high school students. *Motivation Science*, 4(2), 158-176. <https://doi.org/10.1037/mot0000076>
- O'Connor, A. J., Nemeth, C. J., & Akutsu, S. (2013). Consequences of beliefs about the malleability of creativity. *Creativity Research Journal*, 25(2), 155-162. <https://doi.org/10.1080/10400419.2013.783739>
- Pannells, T. C., & Claxton, A. F. (2008). Happiness, creative ideation, and locus of control. *Creativity research journal*, 20(1), 67-71. <https://doi.org/10.1080/10400410701842029>
- Price, L. L., Coulter, R. A., Strizhakova, Y., & Schultz, A. E. (2018). The fresh start mindset: Transforming consumers' lives. *Journal of Consumer Research*, 45, 21-48. <https://doi.org/10.1093/jcr/ucx115>
- Puccio, G. J. (2017). From the dawn of humanity to the 21st century: Creativity as an enduring survival skill. *The Journal of Creative Behavior*, 51, 330-334. <https://doi.org/10.1002/jocb.203>
- Puente-Díaz, R., & Cavazos-Arroyo, J. (2017). The Influence of creative mindsets on achievement goals, enjoyment, creative self-efficacy, and performance among business students. *Thinking Skills and Creativity*, 24, 1-11. <https://doi.org/10.1016/j.tsc.2017.02.007>
- Puente-Díaz, R., & Cavazos-Arroyo, J. (2019). Creative mindsets and their affective and social consequences: A Latent class approach. *The Journal of Creative Behavior*, 53(4), 415-426. <https://doi.org/10.1002/jocb.217>
- Rhew, E., Piro, J. S., Goolkasian, P., & Cosentino, P. (2018). The Effects of a growth mindset on self-efficacy and motivation. *Cogent Education*, 5(1), 1492337. <https://doi.org/10.1080/2331186X.2018.1492337>
- Rissanen, I., Kuusisto, E., Tuominen, M., & Tirri, K. (2019). In search of a growth mindset pedagogy: A Case study of one teacher's classroom practices in a Finnish elementary school. *Teaching and teacher education*, 77, 204-213. <https://doi.org/10.1016/j.tate.2018.10.002>
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1), 1-28. <https://doi.org/10.1037/h0092976>

- Royston, R., & Reiter-Palmon, R. (2019). Creative self-efficacy as mediator between creative mindsets and creative problem-solving. *The Journal of Creative Behavior*, 53(4), 472-481. <https://doi.org/10.1002/jocb.226>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Schunk, D. H., & DiBenedetto, M. K. (2016). Self-efficacy theory in education. *Handbook of Motivation at School*, 2, 34-54.
- Seelig, T. (2015). *Insight out: Get ideas out of your head and into the world*. HarperCollins.
- Sternberg, R. J., & Lubart, T. I. (1999). The Concept of creativity: Prospects and paradigms. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 3-15). Cambridge University Press. <https://doi.org/10.1017/CBO9780511807916.003>
- Stolaki, A., & Economides, A. A. (2018). The Creativity challenge game: An Educational intervention for creativity enhancement with the integration of Information and Communication Technologies (ICTs). *Computers & Education*, 123, 195-211. <https://doi.org/10.1016/j.compedu.2018.05.009>
- Tan, J., Yap, K., & Bhattacharya, J. (2021). What does it take to flow? Investigating links between grit, growth mindset, and flow in musicians. *Music & Science*, 4, 2059204321989529. <https://doi.org/10.1177/2059204321989529>
- Usher, E. L., Li, C. R., Butz, A. R., & Rojas, J. P. (2019). Perseverant grit and self-efficacy: Are both essential for children's academic success? *Journal of Educational Psychology*, 111(5), 877-902. <https://doi.org/10.1037/edu0000324>
- Wang, S., Dai, J., Li, J., Wang, X., Chen, T., Yang, X., He, M., & Gong, Q. (2018). Neuroanatomical correlates of grit: Growth mindset mediates the association between gray matter structure and trait grit in late adolescence. *Human Brain Mapping*, 39(4), 1688-1699. <https://doi.org/10.1002/hbm.23944>
- Weiner, B. (1985). An Attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4), 548-573. <https://doi.org/10.1037/0033-295X.92.4.548>
- White, K., & McCoy, L. P. (2019). Effects of game-based learning on attitude and achievement in elementary mathematics. *Networks: An Online Journal for Teacher Research*, 21(1), Article 5. <https://doi.org/10.4148/2470-6353.1259>
- Yang, K.-H., & Chen, H.-H. (2021). What increases learning retention: Employing the prediction-observation-explanation learning strategy in digital game-based learning. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2021.1944219>
- Yeh, Y. (2017). Research development of creativity. In J. Stein (Ed.), *Reference Module in Neuroscience and Biobehavioral Psychology* (pp. 1-11). Elsevier.
- Yeh, Y. (2020). *Integrating growth mindset into digital creativity game-based learning: Its effects on urban and remote area pupils' creativity mindset and creativity*. Project report (MOST 107-2410-H-004-079-SS2). Ministry of Science and Technology.
- Yeh, Y., & Lin, C. S. (2018). Achievement goals influence mastery experience via two paths in digital creativity games among elementary school students. *Journal of Computer Assisted Learning*, 34(3), 223-232. <https://doi.org/10.1111/jcal.12234>
- Yeh, Y., Chang, H. L., & Chen, S. Y. (2019). Mindful learning: A Mediator of mastery experience during digital creativity game-based learning among elementary school students. *Computers & Education*, 132, 63-75. <https://doi.org/10.1016/j.compedu.2019.01.001>
- Yeh, Y., Sai, N. P., & Chuang, C. H. (2020). Differentiating between the "need" for and the "experience" of self-determination regarding their influence on pupils' learning of creativity through story-based digital games. *International Journal of Human-Computer Interaction*, 36(14), 1368-1378. <https://doi.org/10.1080/10447318.2020.1750793>
- Zander, L., Brouwer, J., Jansen, E., Crayen, C., & Hannover, B. (2018). Academic self-efficacy, growth mindsets, and university students' integration in academic and social support networks. *Learning and Individual Differences*, 62, 98-107. <https://doi.org/10.1016/j.lindif.2018.01.012>
- Zhao, Y., Niu, G., Hou, H., Zeng, G., Xu, L., Peng, K., & Yu, F. (2018). From growth mindset to grit in Chinese schools: The Mediating roles of learning motivations. *Frontiers in Psychology*, 9, 1-7. <https://doi.org/10.3389/fpsyg.2018.02007>
- Zhou, Y., Yang, W., & Bai, X. (2020). Creative mindsets: Scale validation in the Chinese setting and generalization to the real workplace. *Frontiers in Psychology*, 11, 463. <https://doi.org/10.3389/fpsyg.2020.00463>

Appendix A The employed inventories

Table A1. The test items and Cronbach's α of the creativity mindset inventory

Items	Factor loading
Growth Mindset ($\alpha = .911$)	
Factor 1: Growth-Internal locus of control (GI) ($\alpha = .859$)	
9 I can be more creative as long as I am willing to learn.	.855
1 As long as I work hard, my creativity can be greatly improved.	.801
5 I can improve my creative ability through self-learning.	.700
Factor 2: Growth-External locus of control (GE) ($\alpha = .850$)	
2 My creativity can be improved with the help of good teachers.	.849
10 My creativity can be substantially improved when I have sufficient learning opportunities.	.789
6 I am willing to learn creativity and I can become more creative when there is a good learning environment.	.753
Fixed mindset ($\alpha = .952$)	
Factor 3: Fixed-Internal locus of control (FI) ($\alpha = .877$)	
3 It is hard to improve my creativity even if I work hard to improve it through self-learning.	.872
7 Even if I am willing to learn creativity, it is hard for me to become more creative.	.522
11 Even if I work hard by myself, my creativity won't be substantially improved.	.476
Factor 4: Fixed-External locus of control (FE) ($\alpha = .924$)	
12 Even if I have sufficient learning opportunities, my creativity won't be substantially improved.	.909
8 Even if there is someone to tutor me, it's hard for me to become more creative.	.741
4 It is hard to improve my creativity even if I have good luck and meet good teachers.	.409

Note. Sources of construct development: Dweck (2007), Karwowski (2014), and Rotter (1966).

Table A2. The test items and Cronbach's α of the Grit Scale ($\alpha = .872$)

No.	Factor 1: Perseverance of Effort ($\alpha = .872$)	
3	I am diligent.	
4	I am a hard worker.	
1	I finish whatever I begin	
6	Once I am obsessed with a certain idea or project, I won't lose interest.	
	Factor 2: Consistency of Interest ($\alpha = .813$)	
2	Setbacks don't discourage me.	
7	I can maintain my focus on projects that take more than a few months to complete.	
8	New ideas and projects won't distract me from previous ones.	
5	Once I set a goal, I will try to pursue it and won't give up easily.	

Note. Sources of construct development: Duckworth and Quinn (2009) and Duckworth et al. (2007).

Table A3. The test items and Cronbach's α of the Inventory of Self-Determination in Digital Games ($\alpha = .933$)

No.	When playing the game,	
	Factor 1: Autonomy and self-regulation ($\alpha = .887$)	
8	I could freely choose the avatar in the game.	
13	I could freely employ my problem solving strategies.	
12	I had many chances to make free choices.	
3	I could soon forget negative feelings from getting low scores and focus on the next game.	
4	I had abundant opportunities to develop my own thoughts.	
6	I could understand why I failed and immediately adapt to get a higher score.	
9	I could decide the order of game playing	
	Factor 2: Competence ($\alpha = .881$)	
2	I could think of the answer quickly.	
11	I could quickly figure out methods for problem solving.	
1	I performed well.	
7	I could achieve the scores or goals that I set.	
10	I could quickly learn how to achieve high scores.	
5	I felt that the problems or challenges matched my ability level.	

Note. Sources of construct development: Yeh et al. (2019) and Bandura (1977).

Table A4. The test items and Cronbach's α of the inventory of self-efficacy in creativity digital games ($\alpha = .927$)

No	When playing the game,
Factor 1: Ability to generate creative ideas ($\alpha = .908$)	
8	I believe that my creativity can be improved as long as I try hard to learn.
5	I believe that my creativity can be constantly improved.
6	I believe that I can come up with many creative ideas.
4	I believe that I can come up with many creative problem-solving solutions.
7	I believe that I can become a creative person.
9	I believe that I can produce creative works.
Factor 2: Achievement of creative performance ($\alpha = .844$)	
2	I feel that I am more creative than most of my classmates.
1	I feel that I am a creative person.
3	I feel that "being creative" is one of my characteristics.

Note. Sources of construct development: Yeh and Lin (2018), and Ryan and Deci (2000)

Semiotic Alternations with the Yupana Inca Tawa Pukllay in the Gamified Learning of Numbers at a Rural Peruvian School

Rosario Guzman-Jimenez^{1*}, Dhavit-Prem², Alvaro Saldívar² and Alejandro Escotto-Córdova³

¹Universidad de Lima, Perú // ²Asociación Yupanki, Perú // ³UNAM-Zaragoza, México // rguzman@ulima.edu.pe // dhavitprem@gmail.com // yachay@yupanainka.com // aescotto@unam.mx

*Corresponding author

(Submitted February 22, 2022, Revised June 27, 2022, Accepted July 8, 2022)

ABSTRACT: Yupana Inca Tawa Pukllay (YITP) is a ludic didactic resource based on semiotic alternation that, using the reading of numbers in the Inca numeral system, improves its equivalent Indo-Arabic reading. Twelve children from first to fourth grade of a bilingual (Spanish-Quechua), multi-grade elementary school in a small rural Peruvian community were assigned an electronic tablet with YITP and learned autonomously, without teachers during the COVID-19 pandemic. The results obtained show that: (a) they learned in a very short period of time (14 min - 05h 41 min) (b) they improved digit reading accuracy on the first attempt (c) they improved digit reading speed d) they achieved a high percentage of correct reading of numbers containing at least one zero digit. The results suggest the potential of YITP as an educational tool in the teaching-learning process of arithmetic.

Keywords: Semiotic alternations, Yupana Inca Tawa Pukllay, Ethnomathematics gamification, Numeral system, Zero

1. Introduction

Rural education is offered in rural territories and cultures located in different regions of the country that present environmental, linguistic, geographical, historical, and cultural particularities. Schools usually serve very small groups of students (Figueroa et al., 2021) with associated problems such as poverty, isolation, accessibility, available resources, etc.

The Peruvian Ministry of Education (MINEDU, 2021) has established a competency-based model for the area of mathematics, with quantity problem solving as the first competency. However, it maintains the traditional dynamics of classes focused mainly on grades rather than on the teaching-learning process. The didactic materials are maintained without the support of appropriate technology according to the new pedagogical strategies and objectives, and traditional teaching dynamics are maintained. In the results obtained from the PISA tests from 2015 to 2018, a slight improvement is observed but the most basic levels of performance in mathematics are maintained (MINEDU, 2018).

The health emergency caused by the COVID-19 pandemic led the temporary closure of schools in both urban and rural areas, forcing students to interrupt their on-site studies and opt for the emergency solution implemented by the Peruvian government: “Aprendo en casa” (“I Learn at Home”) program, which was launched even without the minimum conditions for adequate distance education, seeking to adapt materials, content and media to the national curriculum for basic education (Andrade & Guerrero, 2021) but without an adequate strategy in the use of educational technology. During the last two years, mainly rural students have not had access to classes in multigrade schools, have lacked the presence of teachers and/or adequate feedback on their progress during the deployment of such government programs.

Migrating from a face-to-face educational system to a distance learning system requires not only the acquisition of materials and media, but mainly the use of innovative and adequate methods and methodologies for this purpose, which if necessary use semiotic alternations that allow a more effective learning process and if possible, accelerate it through the development of educational materials that involve the student, capturing their attention by integrating the cultural, scientific and technological elements related to their genuine interest.

The learning of the numeral system, in the case of the first grades of primary education, is an essential basis for students’ understanding of arithmetic and consequently of mathematical skills in general, as well as for their applications throughout their daily activities. This context raised the following research questions for us:

Given the negative circumstances: (a) 1st and 2nd year children never received previous face-to-face education at school, (b) YITP is a new tool inserted in primary education, (c) students have no digital experience, (d) most of students' parents are illiterate, (e) inexistence of wifi connection in the community and isolation due to the pandemic made difficult hardware and software support; will children of a multi-grade school in the Cañaris community of Peru learn to read Indo-Arabic numbers and their equivalent Inca numbers through the semiotic alternation of Yupana Inca Tawa Pukllay (YITP) embedded in an electronic tablet ?

In the learning process of Indo-Arabic and Inca number systems reading using the digital YITP serious game in tablets, what peculiarities will be shown by:

- 1st and 2nd grade children who never had previous face-to-face arithmetic education and
- 3rd and 4th grade children who did have previous face-to-face arithmetic education at school?

What differences in the learning process of Indo-Arabic and Inca number systems reading using the digital YITP serious game in tablets between 1st and 2nd grade versus 3rd and 4th grade will there be?

The present research was developed in a context of geographic isolation, social interactions, and socioeconomic isolation of the Cañaris community, accentuated by the confinement of epidemiological policies in the face of the COVID-19 pandemic, with the absence of a multigrade teacher to teach the reading of Indo-Arabic numbers in arithmetic for elementary school children. It was a quantitative correlational type research with an epistemological, semiotic and ethnomathematical approach to the learning process of children in the numeral system through the use of the Yupana (Inca abacus) and its arithmetic method called YITP, for which the objectives were: (1) Demonstrate that YITP within the Tablet (SERO-TP) will facilitate learning Indo-Arabic numeracy in children with relative isolation due to the COVID-19 pandemic. (2) To compare different parameters of the learning of one-digit to five-digit numbers reading in two groups of children, group a: 1st and 2nd grade (never assisted classes at school before) and group b: 3rd and 4th grade (who assisted classes at school for 1 or 2 years before pandemics), using the SERO-TP. (3) To provide empirical evidence that the didactic use of semiotic alternations as a resource for mathematical learning and self-learning, in this case, the learning of numbers in the Indo-Arabic system by the non-Indo-Arabic YITP, is an intuitive educational support, playful, which facilitates learning and shortens the time in which arithmetic contents are learned and mastered. (4) To provide empirical evidence that the yupana inca, developed centuries ago in Peru and recently proposed as the YITP method is a didactic resource easy to use by children, and therefore, useful, economical, and easy to implement for the current teaching of arithmetic in primary school, with relative independence of the socioeconomic, linguistic, and cultural condition of the students. (5) To provide empirical evidence for our proposal that the YITP has the symbolic representation of place-valued zero in a visuospatial matrix, and that its visuospatial representation, used as a semiotic alternation, facilitates the mastery of learning Indo-Arabic numbers with place-valued zeros.

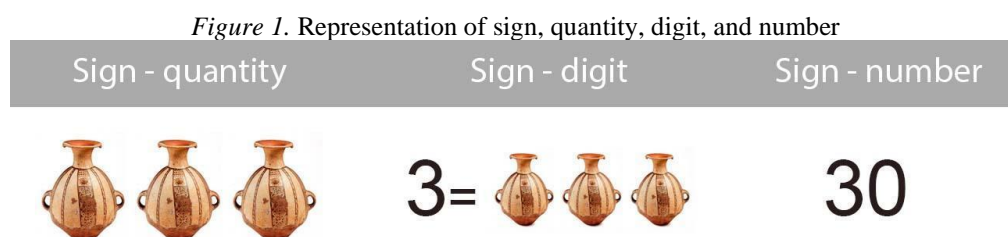
2. Literature review

Semiotic alternations (SA) are the change of one sign for another to represent the same meaning. Their main effect is to make it more precise, or to expand it, or to clarify it (Escotto-Córdova, 2021). SAs enhance thinking, facilitate the generalization of concepts and are a didactic resource used daily in the teaching and learning of any subject of knowledge, and mathematics is no exception. Examples of semiotic alternations are metaphors, drawings, objects, body movements accompanying speech, diagrams, writing, still or moving images in videos, etc.

The key to understanding the use of semiotic alternations as a didactic resource is to be precise in the concepts of sign and meaning that we use. By sign we will understand any physical entity that stands in place of something (a physical entity or a conceptual entity) for someone. The physical form of signs is varied: phonic, gestural, facial, manual or corporal; objectual, wavelengths, non-iconic graphics (e.g., writing), iconic graphics (e.g., drawings). We will understand by meaning everything that is substituted by a sign, be it a physical or conceptual entity. There is no sign without meaning, but they are not the same thing. The same sign can have different meanings, or the same meaning can be expressed by different signs. In mathematics (socioculturally constructed systems of signs and meanings) both conditions occur, particularly when we speak of quantity, digit and number.

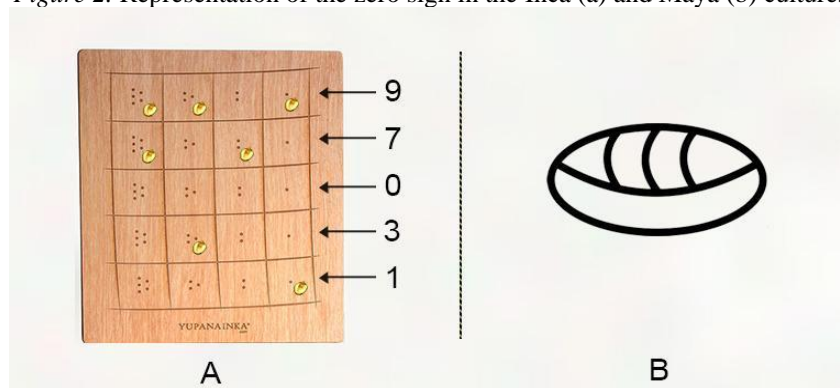
A sign whose meaning is quantities of something is not the same as one whose meaning is conceived as a digit, nor are both equivalent to one whose meaning is conceived as a number. A sign-quantity shows, evidences, points out one by one, a certain quantity of entities; a sign-digit refers to a set of entities whose quantity is

expressed in the sign, it implies an order and a hierarchy with respect to other signs-digits; a sign-number, has its meaning defined by other signs and meanings of a specific mathematical system with rules, for example, the number zero. (Escotto-Córdova, 2021).



In Figure 1 in the first cell, the sign-quantity shows a quantity of objects; in the second, sign-digit, the digit 3 means a set of a quantity of entities, and maintains an order and hierarchy with respect to other digits such as 2 or 4; in the third cell, sign-number, the digit 30 (thirty) carries a three followed by a zero, this sign “zero” in no way means an absent quantity, that is, “no physical entities.” It is not a sign of quantity, but a number sign whose meaning we paraphrase as follows: “sign that refers to the fact that the digit that accompanies it multiplies the quantity it represents the same times by the base of the numeral system,” in this case ten ($3 \times 10 = 30$), and that gives a positional value of the digit in the tens, in the specific example.” That meaning is not given by the sign-digit, much less by the sign-quantity (Escotto-Córdova, 2021). The importance of this theoretical distinction will be clearly seen with the YITP and the representation of the zero on the board and the tokens with which the quantities of the decimal system are represented in the YITP. Its historical significance is that the Incas represented zero in the yupana by leaving an empty row (without tokens), and a space without knots in the khipu (Pereyra, 1990; Prem, 2016; Urton, 2005), this being a specific type of sign for zero (see Figure 2A) different from that used by the Maya in our continent (see Figure 2B).

Figure 2. Representation of the zero sign in the Inca (a) and Maya (b) cultures



The distinction between sign-quantity, sign-digit and sign-number is not an idle one; it implies fundamental conceptual and theoretical changes in the history of mathematics. The use of the sign as a quantity, as a digit or as a number empowers or restricts mathematical thinking, in Table 1 the subject of zero is analyzed.

Table 1. Example of table Meanings of sign-quantity, sign-digit and sign-number

Zero → sign-quantity	Zero → sign-digit	Zero → sign-number
0 (zero)		$3^0 = 1$
It means nothing, nothing, or entity, so that in 30, we understand 3 together with nothing results in three.	Zero means a positional place, so that in the digit 30, we never interpret three and nothing, but a set of thirty entities	Zero does not mean just a positional place, its meaning depends on a system with specific rules. In this case exponents indicate an operation in which its result is one

In general terms, we could say that in quantities, the elements that compose them are perceived one by one. That is why there is a biological limit common to several species of animals, in babies’ weeks old and in cultures and languages called “anumerical.” Its cortical foundation is usually located in the intraparietal sulcus (Chrisomalis, 2004; Dehaene et al., 2003; Dehaene et al., 2001; Everett, 2009; Lupyan & Bergen, 2016; Wassman & Dasen, 1994; Wiese, 2007).

If quantities are perceived, digits and numbers are conceived. The first step is to conceptualize sets of quantities, i.e., the notion of digits. “The Karitian people say, ‘take one’ and show another hand to name the digit six” (Everett, 2019, p. 80), thus used, the hand is a set of a certain number of fingers, i.e., a digit. The next historical step was to advance to the notion of “number,” that is, a sign whose meaning depends on other signs and meanings within a mathematical system with specific rules for its relationships. This step occurred with negative numbers and zero.

In the above examples we have cases in which the same sign has different meanings (as quantity, as digit and as number), which in itself complicates the understanding of arithmetic in children when they understand the digit-signs only as quantities of entities, and not as a set of quantities, or even worse, they confuse numbers with quantities.

Semiotic alternations have been present in every advance and development in the history of mathematics, in all times and in any culture. We can identify three main stages in the creation and use of signs to signify different mathematical concepts: the first, signs with the meaning of quantity: they represent by means of any figure or by means of objects each entity of a specific quantity, for example, to represent five put five points or five fingers. The second, signs with the meaning of digits: they use a figure or object to represent a set of entities, that is, the quantity as a set. For example, they use the hand as a set to represent five entities or fingers, or the human body to represent twenty. And the third, signs with meaning of numbers, are graphic signs or physical entities as engravings, whose meaning is defined by other sign-meanings of a mathematical system, for example, the vigesimal or the decimal, including the operations and relations between the sign-meanings. The clearest example is zero (Cajori, 2011; Escotto-Córdova, 2021; Dehaene, 2016; Ifrah, 2000; Menninger, 1992).

The signs at each stage of mathematical development used by different cultures and at different times have been varied: physical entities as a sign of quantities, for example each of the fingers of the hand, or parts of the body have been used from the representation of quantities with objects or things (Dehaene, 2016; Everett, 2019), or simply lines, drawn sticks. For example, the Ishango bone, from twenty thousand years ago, has marks of quantities representing one by one each counted entity (Everett, 2019). In Egypt of the first millenia B.C., eight is represented by eight lines, sticks or vertical marks; the same in Iran, Elam culture, in the third millennium B.C.; the same in the Indian civilization of 2300-1750 B.C.; or the Hittite civilization, in Anatolia, between 1500-800 B.C.; in Greece between the 5th and 2nd century B.C.; and in the Lydian civilization between the 6th and 4th century B.C. (Ifrah, 2000; Menninger, 1992).

The transition to signs with meaning of a set of quantities is already noticeable with the Theban Greeks or the Chaldeans who began to use a sign to represent the set of five between the 5th and the 1st century B.C.; the same was the case in the Lydia civilization between the 5th and the 4th centuries B.C.; also in Asia Minor, between the first half of the first millennium B.C.; and with Mayans between the 3rd and 4th centuries B.C. (Ifrah, 2000; Menninger, 1992).

Finally, an example of the transition from signs as a set of quantities to signs as numbers occurs with the notion of zero in India, in Central America with the Mayan culture representing zero with a drawing of an empty shell, or the cultures of Cambodia (Khmer culture), Viet-Nam, Laos with the representation of a point (Aczel, 2015).

3. Methods and materials

3.1. Tawa Pukllay Method in the Yupana (YITP)

Those who only focus on its instrumental function, overlook the semiotic nature (signs and meanings) of the YITP method and the yupana device itself, and consequently fail to recognize the power that signs and meanings carry as a “cognitive tool” (Lupyan & Bergen, 2016; Vygotski, 2017), particularly evident in numbers (Everett, 2019). In terms of the cognitive work involved in the use of YITP, it has the virtue of decreasing the working memory load with respect to that required in the Indo-Arabic arithmetic system, since YITP notation and operations depend on visuospatial relationships, pattern recognition, simple movements and a full-time visualization of quantities and performed operations. In theoretical-conceptual terms, it is useful for the understanding of the arithmetic system including zero as a number: YITP facilitates the distinction between sign-quantity, sign-digit, and sign-number.

According to Prem (2018a) the YITP is a board with squares arranged in 4 columns and 5 rows in which any number up to 5 digits can be represented: 0 to 99999. If it is required to work with larger quantities, rows can be

added upwards representing the powers of 10^5 , 10^6 , 10^7 , etc. In each row the cells are marked from left to right with 5, 3, 2 and 1 dots. The numbers are represented by placing small objects such as seeds (tokens) in the cells. Each row in the board represents a digit. It starts with the bottom row to represent the digit for units; the next row up for the tens digit and so on, up to the ten thousand. According to this rule, the digits 1, 2, 3 and 5 are represented with a single token with quantity-sign function (one entity), which is placed in the cell with the dots engraved on it. This is to represent the sign-digit of one, two, three and five entities. The cell is the sign conformed by a set of dots whose rule indicates that it can be used in writing a number by placing only one token on it or none at all. For example, the digit 7 is represented with one token in cell 5 and one token in cell 2. Each row of the board represents a positional value from bottom to top with the meaning of units, tens, hundreds, etc., for example, the number 47, with a token in cells 5 and 2 in the bottom row and the next row up with a token in cells 3 and 1 (see Figure 2A).

With these elements we can already distinguish signs-quantities, signs-digits, and signs-numbers. The tokens from right to left and from top to bottom represent the order and hierarchy of the digits, and that their meaning is numerical, since they depend on the YITP system of digit-signs with exponential multiples with base 10. Now, since in each cell one or no token is placed to indicate quantities (the token is not counted, but indicates the number of points indicated in that cell), they only signify quantities, which located in one or more cells of a row form a group of digit cells of the YITP. That is, each row represents a digit that in turn, depending on its order and the rules of the numeral system (base 10 and exponential multiples), will form part of a number-sign.

Zero, in its double condition of nothing and number is clearly evident when we have a multi-digit number, for example 97031 (see Figure 2A), which contains representations of digits that are arranged in five rows of the YITP, having in the top row the representation of nine, then the representation of seven, no token in the third row, and in the rows below are represented three and one respectively. All of the above decreases the working memory load and facilitates the theoretical-conceptual understanding of the notions of “number,” “digit,” “quantity” and the importance of zero in a certain place value.

The semiotic properties of the YITP, both numerical and didactic, can be a valuable tool for the reading of numbers in elementary school children both in the city and in rural places, but above all, in those who due to their socioeconomic, cultural and social conditions have not had more academic support than that provided in their schools by their teachers and friends. Therefore, we proposed to provide evidence of the sociocultural usefulness for rural education in a context of pandemic, which improves the speed and effectiveness in reading the inca numeral system in the YITP and its equivalence to the Indo-Arabic system through a self-learning serious game on an electronic tablet used by the children of the community.

3.2. Serious Game SER0-TP

A serious game is an educational application whose main purpose is to coherently combine serious aspects such as teaching, learning, communication or even information with fun aspects of video games in a non-exclusive, non-exhaustive way (Alvarez, 2007).

Education researchers have taken a keen interest in gamification since 2013 (Dominguez et al., 2013). The gamification in education is an intense and quickly developing area of research, with hundreds of new relevant publications coming out every year (Lee & Hammer, 2011). Gamification has also been shown to have favorable results, relating its use to greater student engagement and learning (Tsai et al., 2019; Díez et al., 2017)

Studies show that gamification can make a positive contribution to the education process (Kim & Castelli, 2021; Manzano-Leon et al., 2021; Swacha, 2021). Serious Game is a wide field that may be used for many educational purposes. Since it is also a mean of entertainment, multiple learning objectives can be covered while many skills are developed at same time: information technology, communication, language and actually almost any field. And one very important thing, specially committed to fulfill the multigrade schools' needs: It is for all ages. (Mouaheb et al., 2012). Learners with different skills can participate effectively in the same learning application (Sezgin et al., 2018) and the more they get engaged, the more they understand their own learning process. This process stimulates student's autonomy at learning time in a more effective way (Lee & Hammer, 2011).

The SER0-TP serious game installed on an electronic tablet was designed following the learning guidelines proposed by the sociocultural theory of learning and development (Vygotski, 1995; Vygotski, 2010; Vygotski, 2017), Galperin's theses with his theory of knowledge formation by stages (Galperin, 2009a; Galperin, 2009b; Galperin, 2009c; Talizina, 2009), and Leontiev's activity theory (Leontiev, 1984; Leontiev, 2009). These guidelines consist in the fact that psychological development and learning are always carried out with the help of

others, by others, for others until the moment arrives when the individual (child or adult) does it for himself as if he were someone else using his internal language. Its practical expression is self-learning. Therefore, learning takes place in stages, some external (social, object, figurative-drawings, etc.) and other internal (oral language: one speaks when learning; and another silent: one speaks to oneself silently). In the research, the external aspect was the initial orientation of the teacher and the tutorials, the object aspect was the electronic tablets, and the internal and playful aspect was the game played by each child at his own pace and taste as a manifestation of his self-learning.

The design of the mechanics, dynamics and aesthetics, the guidelines proposed for MDA (Hunicke et al., 2004) were used. Figure 3 shows Module 1: PUKLLAY, which contains the interactive exercises, as well as the children in the middle of learning.

The tablet was programmed to record frequencies of use for each level, successes, times, etc., whose numerical data were later analyzed and statistically processed.

3.3. Participants

The research was carried out in the community of Huamachuco, where there are approximately 25 families (25 fathers, 25 mothers and an average of between two and three children per family), for a total of approximately 125 inhabitants, most of whom work in rural agriculture and are illiterate. The language spoken by the majority is Quechua (Cañaris variant). The families live in single room houses, separated in a distance around 10 to 20 minutes by walking from each other (there is no car or bus transportation within the community, except for a few motorcycles that are used for very specific purposes). The twenty children have their fathers and mothers at home. Twelve children (4 boys and 8 girls) from the first four grades of primary school enrolled in the educational institution N° 10244 multigrade of the community of Huamachuco (Peru), participated in the project. The learning process lasted approximately two months.

The selection of the sample was intentional and exhaustive: all the children of the community who were present and agreed to participate were divided into two groups: those children of 1st and 2nd grades, who never had had any classroom learning experience as a result of the quarantine due to the COVID-19 pandemic and children of 3rd and 4th grades who already had knowledge of the Indo-Arabic numeral system because of their past two years studying at regular school, see Table 2.

Table 2. Boys and girls by academic grade and age

Grade	Boys (age)	Girls (age)	Total
1°		1A,1B (6 years)	2
2°		2A,2B,2C (7 years)	3
3°	3B,3C (8 years)	3A (8 years)	3
4°	4B, 4D (9 years)	4A, 4C (9 years)	4

3.4. Experimental design

The experimental design, as shown in Table 3, worked with 3 types of variables: (1) the independent manipulated variable (IMV), which is a variable that can be manipulated in its magnitudes, frequency, the sequence of its presentation, the time in which it is presented and how long it lasts, it can be placed and removed at the experimenter's will in each subject or group, etc. For the research we considered the YITP programmed on an electronic tablet under the criteria of a serious game which we have called SER0-TP. (2) the independent variable of selection (IVS), which is any variable whose only possible manipulation is to select that it is present or absent in a group, we have considered the Self-learning Group without previous presential learning experience (1st - 2nd grade) and the Self-learning Group with previous presential learning experience (3rd - 4th grade). And finally the following were considered as dependent variables to measure the learning process: (3) Digit reading speed rate (DRSR): digit reading speed measured in seconds per digit, Zero read attempt ratio (ZRAR).

The empirical work of the Tawa Pukllay method has been disseminated by Asociación Yupanki (Dhavit Prem and Alvaro Saldívar) through workshops aimed at students and teachers at both city and community levels in Peru and has been presented in several international events in Colombia (Saldívar et al., 2019a), México (Saldívar et al., 2019b) and Peru (Saldívar, 2019). Also some other events are National Council of Science, Technology and Technological Innovation of Perú (CONCYTEC) 2015; National Library of Perú (BPN) 2017, National Institute of Peruvian Culture INC-Cusco, 2016; Science & Engineering Festival in Washington DC

2017; Latinoamerican Congress of Mathematics (RELME) 2017 and 2018; VI International Congress of Ethnomathematics 2018; High School Academy Congress Guatemala 2017.

Table 3. Learning parameters

Group	Independent variable of selection (IVS))	Independent manipulation variable (IMV)	Dependent variables
Group 1	Self-learning Group without presential learning experience (1 st - 2 nd grade)	SER0-TP	FAAR DRSR ZRAR
Group 2	Self-learning Group with past presential learning experience (3 rd - 4 th grade)	SER0-TP	FAAR DRSR ZRAR

Our research was planned to provide theoretical and empirical systematization, as well as support from educational institutions in Peru (Universidad de Lima) with the participation of researchers from Mexico (Facultad de Estudios Superiores Zaragoza, Universidad Nacional Autónoma de México), for the rescue of the YUPANA INCA as a didactic resource for the teaching of mathematics at the elementary level. The research involves the learning of technology inaccessible to children in the community. On the other hand, it has been proposed to continue the support to this community, but in a face-to-face way, when the restrictions of the pandemic or COVID-19 are lifted.

The community was visited, the research proposal was presented and a discussion was held with the adults seeking approval for the participation of their children in the research in a community session with the entire community. The facilitator was the multigrade teacher of Educational Institution 10244. Once accepted, their informed consent about the research was confirmed by the inhabitants of the community of Huamachuco of the Cañaris community, the parents gave their informed consent forms signed and received a SER0-TP kit (electronic tablet, serious game and solar charger), YITP kits (book, physical yupana and tokens) and instructions for use. Biosafety protocols were observed during the pandemic.

Before starting the experimentation, in order to discard cognitive difficulties, each child was psychologically evaluated by the specialist, Dr. Alejandro Escotto-Córdova through the review of previous videos of the interviews to the children and the application of a brief cognitive assessment designed for the andean population. This assessment had as reference the Montreal Cognitive Assessment - MoCa test. Also, our research team designed a test called “Yupay Tupay - Sami” of attitude-emotion towards mathematics based on the Attitudes Toward Mathematics Inventory Test (ATMI) and the representation of responses on a Likert scale with five emotions (very sad, sad, indifferent, happy and very happy) (see Appendix 1) which was applied in order to know the initial and final perception (attitude) of all students towards mathematics. Finally, all children received at least two teacher-guided sessions on how to use the tablet and an introduction on how to play with the SER0-TP game.

No child presented cognitive dysfunction, despite the fact that some were below the norm. This is explained by specialist Dr. Alejandro Escotto-Cordova because being outside the statistical norm, being statistically abnormal, does not necessarily imply being disordered, dysfunctional, sick, suffering from some pathology or developmental incompatibility. It is simply not being like the others to whom the individual is compared. Certainly, any disorder, cognitive dysfunction, pathology or developmental incompatibility implies being outside the norm, but not the other way around. The only data provided by applied statistically standardized tests, is how close or far the individual is from the statistical norm derived from the population sample. Nothing more. Measuring is not diagnosing (Escotto-Córdova et al., 2021). The diagnosis of cognitive dysfunction is suspected when the individual who is below the statistical norm in a test, also presents notable difficulties in learning with the help of others and with new didactic strategies. This was not the case in any of the Cañaris children, and they even learned to read Indo-Arabic numbers only by playing with the electronic tablet and the YITP.

Once all the above was finished, the children were not visited again in their community until after two months in which the data from the tablets were collected, so the children played freely with the tablet as long and as often as they wanted. Remote monitoring was available based on simple phone calls to their teacher for any eventual technical problem support. Only one error occurred due to forgetting the password to access the electronic tablet but it was solved remotely. The accuracy of answers within the game, the scores, time and frequency of use of the electronic tablet was automatically recorded by SER0-TP without the children being aware of it.

Figure 3. SER0-TP Pukllay module and children learning



The experiment consisted of:

- The children had to watch some videos included inside the SER0-TP, where lessons on how to read and write numbers on the yupana board were detailed explained.
- After watching the videos, the children could enter the game, where numbers represented in the yupana board using the YITP method (inca numerical system), should be read, interpreted and written using the Indo-Arabic numerical system by console (see Figure 3). A total of 60 exercises had to be solved: 12 exercises with 1 digit, 12 with 2 digits, and so on up to 12 exercises with 5 digits.
- After solving the 60 exercises, a *congratulations* message would appear saying that the task is over.
- Majority of children found out that by reentering the game and by canceling the end message, it was possible to continue playing it, so they decided to do it, performing many unexpected additional exercises, which also were recorded by the system and now are part of the analysis under the label *UAE*.

External factors to consider:

- Children could not be helped by their parents since most of them are illiterate, much less they knew how to deal with any electronic device.
- The teacher taught them the basic principles for taking care of the tablets (turn on, turn off, entering the SER0-TP, watching videos, running the game and loading batteries). No further lessons were imparted during the whole experimental process.
- In order to make the application SER0-TP more attractive, familiar and understandable to children, the whole design considered pictures of the children themselves, text and voice feedback messages in local quechua language and characters such as avatars, dresses and other signs based in ancient local cosmovision.
- One of the most important points considered at the moment of designing the current experiment, was the reincorporation of the YITP method, which is a recent proposal of rediscovery of the inca's math after 500 years, which uses the inca board for calculations, the inca numerical system, pattern recognition, andean principles and quechua names for token movements (Prem, 2018a). It means a totally different way of reading numbers and performing arithmetical operations than Indo-Arabic classic method currently used worldwide.
- After the experimental process, the teacher extracted the files containing the hidden records (XML files) of exercises performances, times and scores and sent them back to the central in Lima for analysis.
- The teacher also interviewed the children who said that those who had siblings at home or friends living near, could help each other on learning how to use the buttons, enter the videos and the game. They also said that each child did his own exercises because they wanted to. No pressure of time existed, nor regular basis of practice was imposed. There was only a general and open suggestion of "it would be good if practice would occur half an hour a day until you finish the exercises." This suggestion was exceeded because all the children said they enjoyed learning in game mode.

4. Results

At the end of the research, the data recorded from the interaction between the children and SER0-TP was collected for the analysis of the following metrics which include two scenarios: expected exercises scenario (EE) and additional unexpected exercises scenario (UAE), which is a series of exercises that children decided by themselves to solve even after finishing the expected tasks (more exercises than requested).

- First attempt accuracy ratio (FAAR): percentage of number reading accuracy at the first attempt.

- Digit reading speed rate (DRSR): digit reading speed measured in seconds per digit (includes the seconds of reading a digit written in the Inca numeral system and its equivalent writing in the Indo-Arabic system using the SER0-TP). This rate is important because it allows us to observe the learning process of the children and the internalization of the value of the squares that gradually replaced the continuous counting of dots.
- Zero read attempt ratio (ZRAR): percentage of accuracy in reading numbers containing at least one zero as a digit at the first attempt.

The FAAR values curves of the reading numbers exercises containing between 1 to 5 digits show that some of them are ascending and others sinuous during the learning process, in which almost all the children achieved above the 90% FAAR, except for the three 3rd grade boys and one 2nd grade girl. However, these four children also showed upward learning, particularly boy 3B who starts with 8% FAAR and ends with 75% FAAR. Child 3A starts with 75% FAAR at 1-digit reading, and then she gets a 50% FAAR at 2 or more digits reading, apparently due to difficulties when identifying the positional value of the represented digits.

From the results obtained it is worth noting that 1st and 2nd grade children learned the Inca numeral system by reading in an autonomous and playful way and its equivalence to the Indo-Arabic system through the SER0-TP in a range of time <20'11" – 2h 26' 28">, also showing a FAAR increase above the 90% for children who also solved the unexpected additional exercises UAE. The range of SER0-TP usage time of all children is <13'44" – 5h 41'37"> (see Table 4).

It is worth noting that although child 2C started with 0% FAAR at 1-digit reading, he increased his FAAR noticeably, completing all the EE and even performing unexpected additional exercises UAE, where he reached a 91% FAAR. The total SER0-TP usage for this child was around 2.5 hours (see Table 3). It is important to highlight that children 2C, 3B, 4B and 4D, started with the lowest FAAR (0%, 8%, 17% and 25%) and ended with FAAR: 91%, 75%, 100% and 90%, respectively.

Table 4. FAAR per digit and total times of SER0-TP usage (EE & UAE)

Expected Exercises (EE)								Unexpected Additional Exercises (UAE)		
Student	Qty	Time EE	1-digit FAAR	2digits FAAR	3digits FAAR	4digits FAAR	5digits FAAR	Qty	Time UAE	FAAR
1 st Grade										
1A (♀)	49	33'39"	58%	83%	92%	83%	100%	0	---	---
1B (♀)	60	19'17"	75%	83%	83%	92%	75%	294	1h 43'20"	92%
2 nd Grade										
2A (♀)	60	11'25"	100%	83%	100%	100%	92%	25	8'46"	92%
2B (♀)	60	34'43"	75%	100%	92%	75%	83%	0	---	---
2C (♀)	60	31'50"	0%	42%	83%	83%	58%	319	1h 54'38"	91%
3 rd Grade										
3A (♀)	60	54'04"	75%	50%	58%	42%	67%	203	1h 55'27"	74%
3B (♂)	40	13'44"	8%	58%	83%	75%	75%	0	---	---
3C (♂)	60	22'03"	67%	67%	92%	75%	75%	8	4'08"	71%
4 th Grade										
4A (♀)	60	28'02"	92%	92%	83%	92%	67%	813	5h 13'35"	90%
4B (♂)	60	21'35"	17%	92%	83%	100%	92%	4	1'19"	100%
4C (♀)	57	25'55"	100%	100%	58%	75%	100%	0	---	---
4D (♂)	60	24'31"	25%	100%	92%	75%	100%	407	2h 52'45"	90%

Note. The times shown do not consider the minutes spent watching the video tutorials; only the effective time of the exercises they performed was counted.

4.1. Digit reading speed rate (DRSR)

At the suggestion of the authors of the YITP method and methodology, with 8 years of teaching experience, three time ranges were considered: < 0-5]s, < 5-10]s and < 10+>s. The first range corresponds to the DRSR of subitizing (Cheeseman et al., 2021), i.e., children do not need to count and only seeing the marked cells they recognize the represented digit, so it is very fast and constitutes an excellent resource of semiotic alternation to introduce Indo-Arabic arithmetic; the second range corresponds to the counting DRSR (children may already be subitizing in some cases, but they still need to count dots, so reading is slower); the third range corresponds to

the counting DRSR with difficulties and/or typing errors when entering the result, which imply an increase in time because the number needs to be entered again in the answer (see Figure 4 and 5).

Figure 4. Digit reading speed rate (DRSR) for 1st and 2nd grade - Expected exercises

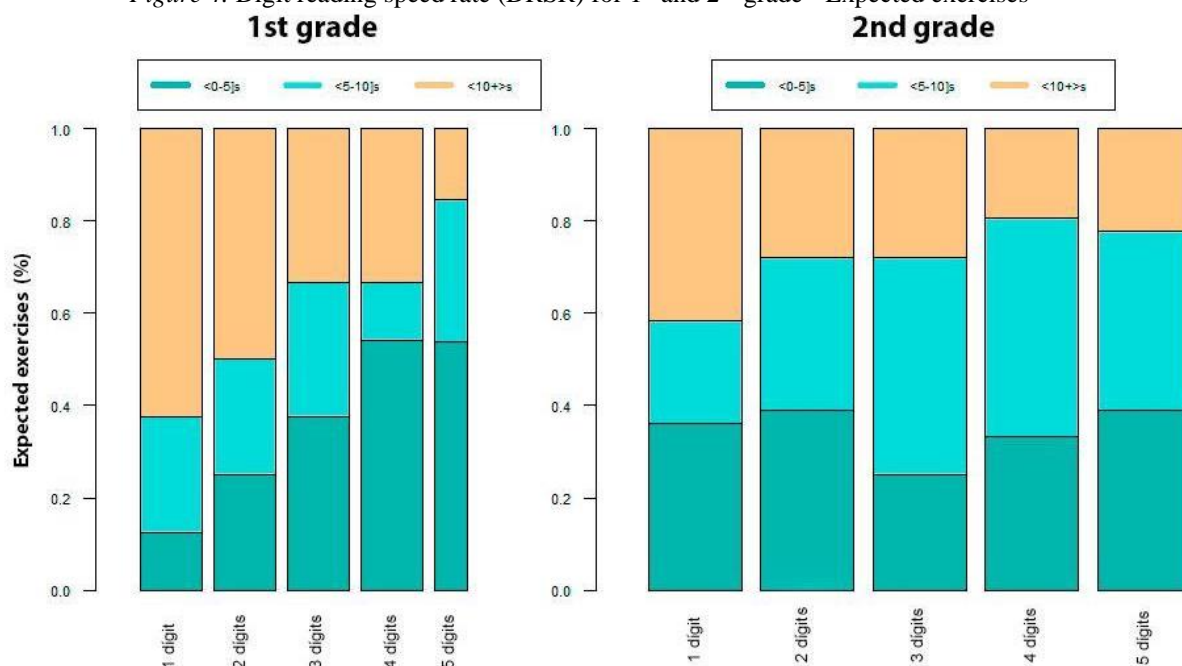
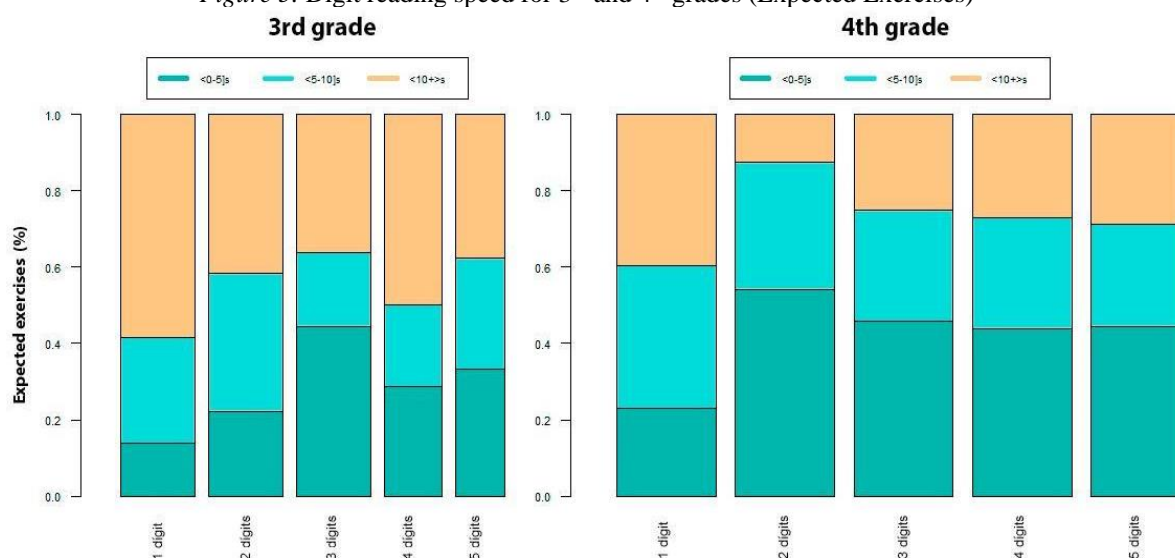


Figure 5. Digit reading speed for 3rd and 4th grades (Expected Exercises)



Note. Range < 0-5]s in dark green, < 5-10]s in light green, and range of < 10+]s in light orange. The thickness of the bar indicates proportionally the number of exercises performed per number of digits.

In all grades it is observed that there is a growth in the exercises solved in the range < 0-5]s, while in contrast, a decrease is seen in the range < 10+]s.

Comparing Figure 6 of UAE with the previous EE graphs (Figures 4 and 5), it is observed that all grades increase the DRSR in the < 0-5] interval, exceeding 70%, with the exception of 3rd grade which rises from 33.33% to 52.15% of DRSR in the < 0-5]s interval.

Figure 6. Digit reading speed rate - Additional exercises

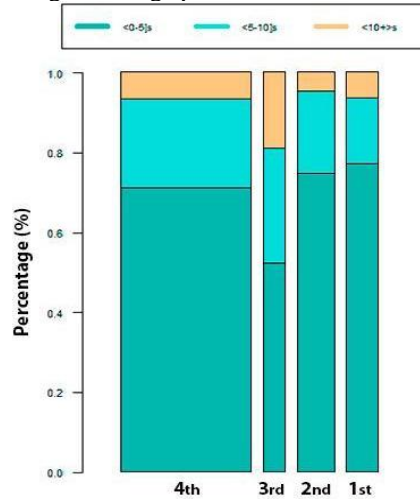
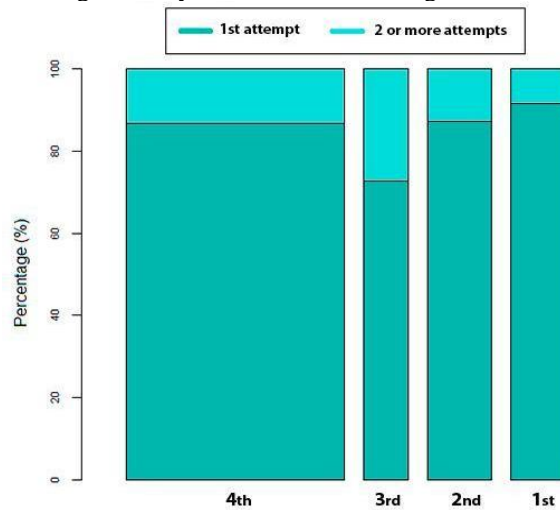


Figure 7. Reading accuracy of numbers containing at least one zero as a digit



Note. Numbers that were read correctly at the first attempt are shown in dark green, while those that required two or more attempts are shown in light green. The thickness of the bars indicates respectively and proportionally the total number of these exercises, which vary by grade because they were generated randomly.

4.2. Percentage in reading numbers containing at least one zero as a digit at the first attempt (ZRAR)

To determine the reading accuracy of numbers containing at least one zero as a digit at the first attempt, the expected EE and additional UAE exercises were included, see Figure 7.

It is observed that the percentage of numbers containing at least one zero as a digit read correctly at the first attempt exceeds 85% for all grades, with the exception of 3rd grade, which reaches 72.73%. The grades with the highest percentages are 1st and 2nd grade with 91.67% and 87.32% respectively.

5. Discussion

Our research suggests that children in a remote community, with educational and technological backwardness, in conditions of temporary isolation due to the pandemic, without a teacher, and with only two months of self-generated play activity, have no difficulty learning two new contents: (1) to use modern technology and, (2) to learn the logic of Inca mathematics installed in a tablet. YITP was shown to be a powerful semiotic alternation that embedded in an electronic tablet (SER0-TP) fosters effective and fast learning, based on a scheme that stimulates playful self-learning.

There is no significant difference when comparing 1st and 2nd grades versus 3rd and 4th grades FAAR. This would suggest that it is not a requirement to have previous basic arithmetical knowledge in order to learn the YITP using the SER0-TP, actually it can be also observed that children from 1st and 2nd grades have achieved, at the end, better FAAR than children from 3rd grade. The similar DRSR of 1st and 2nd versus 3rd and 4th grades and the higher DRSR growth which corresponds to 1st grade, would indicate that previous knowledge of Indo-Arabic system of numbers could demand more time when learning YITP because of cognitive reorganization. ZRAR is high for all four grades and there is no significant difference when comparing the 1st and 2nd grades versus 3rd and 4th grades. This would suggest that YITP would be useful when reading numbers which contain at least one zero.

We argue that in the Quechua culture, in Peru, zero was also represented in the Inca Yupana, but as an absence of quantity in a visuospatial matrix (Prem, 2018b). This representation of zero is intuitive by perceiving the absence of quantity together with its positional value in space, which facilitates performing operations on quantities and decimal-based numbers. Our research suggests that these intuitive properties of the YUPANA INCA for understanding zero are shown in the rapid identification of Indo-Arabic numbers with zeros from using the YUPANA INCA on the electronic tablet, for example 85% of reading numbers containing at least one zero as a digit were read correctly on the first attempt for all grades.

The reading of numbers containing at least one zero as a digit deserves special analysis, since in the Indo-Arabic system it represents a particular challenge for students to distinguish zero with quantitative significance due to its positional value. Nevertheless, in the present investigation a high percentage of reading accuracy was observed on the first attempt of numbers containing at least one digit zero (ZRAR).

The YUPANA INCA used as a didactic resource meets the requirements of the historical-cultural model of the learning process developed by Vygotski et al. (Galperin, 2009a; Galperin, 2009b; Galperin, 2009c; Vygotski, 1995; Vygotski, 2010; Vygotski, 2017; Leontiev, 1984; Leontiev, 2009; Talizina, 2009). based on the transition from object representation to symbolic representation to the internalization and mastery of knowledge, and a few brief explanations and test exercises are sufficient for it to be used playfully as self-learning. Our results provide evidence that this process facilitated by the YUPANA INCA, from:

- Translating quantities into expressions: SER0-TP consists precisely in counting points in the Inca numeral system to be then represented by Indo-Arabic numbers. For this purpose, SER0-TP makes use of the matrix structure of the YITP (see Figure 2), which allows counting dots, adding the quantities of dots from other cells and even accelerating the perceptual process of suddenly recognizing quantities of no more than five objects.
- Using calculation strategies and procedures: SER0-TP allows to easily develop divergent thinking and the associative arithmetic property from the very learning of counting, for example when 5 and 3 dots cells are activated together to represent the number eight.
- Create numerical relationships: SER0-TP stimulates children in the recognition and differentiation of quantities represented in multiple rows, their corresponding positional values and their equivalence in the Indo-Arabic system and facilitates the understanding of zero by showing the sign “empty row” when a power of ten is represented.
- Identifying movements and locating patterns: SER0-TP uses the logic of YITP, which is precisely a proposal for solving arithmetic operations based on the execution of token pattern recognition and the use of a set of moves that depend on the location of these tokens. It is important to highlight that SER0-TP has allowed children in the 1st and 2nd grades of primary school, who in the context of the pandemic did not have the presence of a teacher or tutor for a year and a half, to learn autonomously the numeration of up to five digits. SER0-TP appeals to the autonomy of each child to decide how far to advance, regardless of the grade level, unlike the proposal of current curricula that encourage collective learning.

Our findings show that children have achieved in some cases better levels of digit reading speed at the first attempt RSDR and the percentage of digit reading accuracy at the first attempt FAAR than other children of higher grade, which would demonstrate that with this methodology it is not necessary to limit the learning of number reading to a certain number of digits, and that on the contrary, YITP improves children’s reading accuracy and speed the more digits they have and the greater the number of exercises they solve, as opposed to what is suggested by the official curriculum that promotes learning by segmenting the number of digits that should be learned according to each school grade: 1st grade up to the number 20, 2nd grade up to the number 20, 3rd grade up to three digits, 4th grade up to four digits and 5th and 6th grades up to six digits.

The development of autonomy in students has been highlighted as a goal of mathematics education (Ben-Zvi & Sford, 2007; Yackel & Cobb, 1996). Learner-centered teaching strategies such as mathematics instruction based on real-life contexts, inquiry-based learning, and problem-centered learning (Cobb & Yackel, 1996; Wheatley, 1994) have been discussed to increase learner autonomy in mathematics learning. Intellectual autonomy has been defined as, “students’ awareness and willingness to draw on their own intellectual capacities when making mathematical decisions and judgments” (Cobb & Yackel, 1998, p. 170). Given the pandemic situation, this ceased to be a proposal and became a necessity, a necessity that due to the results obtained shows to be feasible and at the same time a new opportunity that promotes active self-learning.

The results of the adapted attitude test showed an individual improvement, so it is presumed that these mood conditions together with the children’s enthusiasm and curiosity favored self-motivation, which added to the acceptance of the SER0-TP semiotic ludic instrument, allowed the children to achieve the results obtained in a short time.

6. Conclusions

The parameters used in the measurement (FAAR, DRSR and ZRAR) indicate that children in 1st and 2nd grade using the YITP as a didactic instrument learned to read Indo-Arabic numbers. We recommend that the YITP be used as a didactic support in the teaching of arithmetic in the early years in both rural and urban schools. Children in the early grades, even if they have no experience with the Indo-Arabic system, can learn to read using the Yupana YITP as a semiotic alternation. Young children who have not previously had access to technology and who live in isolated conditions can learn to handle technology with playful software in a short time.

The present investigation considers only children from 1st to 4th grade of elementary school; it is suggested to carry out investigations in urban schools and in other rural areas to contrast the results obtained in the present investigation. Although we had a teacher who facilitated the delivery of the SER0-TP kit, and helped in the realization of the cognitive evaluations and initial tests, it would be ideal for specialists in the psychopedagogical area as well as YITP to have direct access to children in future investigations, a fact that could not be done on this occasion due to the limitations of the pandemic.

Rural single-teacher requires didactic strategies and tools to support the individual needs of each student to learn number sense and the notion of place value. It is essential to develop these initial skills that will mark the long-term educational trajectories of these children in mathematics.

Although the present study is only focused on quantity recognition, learning to read numbers in the Inca numeral system and its equivalence to the Indo-Arabic system, it constitutes the basis for a second serious game included in the SER0-TP that will be discussed in a following article.

Acknowledgement

We thank to Scientific Research Center of University of Lima (IDIC), families of Huamachuco Community (Cañaris), teacher Eloy Reyes Huamán and Asociación Yupanki for all the support along this research.

References



















































- Aczel, A. D. (2015). *En busca del cero. la odisea de un matemático para revelar el origen de los números* [Finding zero: A Mathematician’s odyssey to uncover the origins of numbers]. Library Buridan.
- Alvarez, J. (2007). *Du jeu vidéo au serious game: Approches culturelle, pragmatique et formelle* [From video game to serious game: Cultural, pragmatic and formal approaches] (Unpublished doctoral dissertation). Communication and Information Science Speciality, Université Toulouse, France.
- Andrade, A. P., & Guerrero, L. P. (2021). *Aprendo en Casa: Balance y recomendaciones* [Learning at home: Assessment and recommendations]. Grupo de Análisis para el desarrollo. <https://hdl.handle.net/20.500.12820/674>
- Ben-Zvi, D., & Sford, A. (2007). Ariadne’s thread, daedalus’ wings and the learners autonomy. *Éducation et didactique*, 1-3, 117-134. <https://doi.org/10.4000/educationdidactique.241>
- Cajori, F. (2011). *A History of mathematical notations, two volumes in one*. Cosimo Classics.

- Cheeseman, J., Downton, A., Roche, A., & Ferguson, S. (2021). Seeing group structures in subitized numbers. *Broadening Experiences in Elementary School Mathematics*, 7, 147-157.
- Chrisomalis, S. (2004). A Cognitive typology for numerical notation. *Cambridge Archaeological Journal*, 4, 37-52.
- Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*, 31(3-4), 175-190.
- Cobb, P., & Yackel, E. (1998). A Constructivist perspective on the culture of the mathematics classroom. In *The Culture of the Mathematics Classroom* (pp. 158-190). Cambridge University Press.
- Dehaene, S. (2016). *El cerebro matemático: Cómo nacen, viven y a veces mueren los números en nuestra mente* [Mathematical Brain: how numbers are born, live and, sometimes they die in our mind]. Siglo Veintiuno Editores.
- Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, 20, 487-506.
- Dehaene, S., Rivière, D., & LeBihan, D. (2001). Modulation of parietal activation by semantic distance in a number comparison task. *Neuroimage*, 14, 1013-1020.
- Díez, J.C., Bañeres, D., & Serra, M. (2017). Experiencia de gamificación en secundaria en el aprendizaje de sistemas digitales [Gamification experience in secondary education on learning of digital systems]. *Education in the Knowledge Society*, 18(2), 85-105. <https://doi.org/10.14201/eks201718285105>
- Dominguez, A., Saenz-de-Navarrete, J., de-Marcos, L., Fernandez-Sanz, L., Pages, C., & Martinez-Herraiz, J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63, 380-392. <https://doi.org/10.1016/j.compedu.2012.12.020>
- Escotto-Córdova, E. A. (Ed.) (2021). *Alternancias semióticas: estrategia didáctica en la enseñanza de las matemáticas. La enseñanza que aporta la historia de las matemáticas* [Semiotic alternations: didactic strategy in mathematic's teaching. The teaching that supports the history of mathematics]. Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Zaragoza.
- Everett, C. (2009). A Closer look at a supposedly anumeric language. *International Journal of American Linguistic*, 78, 575-590.
- Everett, C. (2019). *Los números nos hicieron como somos*. Crítica.
- Figueroa, T. A., Castro, J. M., Calderon, A. I., & Alburqueque, C. A. (2021). Escuelas rurales en el Perú: Factores que acentúan las brechas digitales en tiempos de pandemia (COVID-19) y recomendaciones para reducirlas [Rural schools in Peru: Factors that increase digital divides in times of pandemics (COVID-19)]. *Educación*, 30(58), 11-33. <https://doi.org/10.18800/educacion.202101.001>
- Galperin, P. Y. (2009a). Acerca del lenguaje interno [About inner language]. In L. Quintanar & Y. Solovieva (Eds.), *Las funciones psicológicas en el desarrollo del niño* (pp. 91-97). Trillas.
- Galperin, P. Y. (2009b). La formación de los conceptos y las acciones mentales [Development of concepts and mental actions]. In L. Quintanar, & Y. Solovieva (Eds.), *Las funciones psicológicas en el desarrollo del niño* (pp. 80-90). Trillas.
- Galperin, P. Y. (2009c). La formación de las imágenes sensoriales y los conceptos [Development of sensorial images and concepts]. In L. Quintanar & Y. Solovieva (Eds.), *Las funciones psicológicas en el desarrollo del niño* (pp. 64-75). Trillas.
- Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A Formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI*, 4(1), 1722. <https://www.aaai.org/Papers/Workshops/2004/WS-04-04/WS04-04-001.pdf>
- Ifrah, G. (2000). *The Universal history of numbers*. John Wiley & Sons.
- Kim, J., & Castelli, D. M. (2021). Effects of gamification on behavioral change in education: A Meta-analysis. *International Journal of Environmental Research and Public Health*, 18(7), 3550. <http://dx.doi.org/10.3390/ijerph18073550>
- Lee, J. J., & Hammer, J. (2011). Gamification in education: What, how, why bother? *Academic Exchange Quarterly*, 15(2), 146-151.
- Leontiev, A. N. (1984). *Actividad, conciencia y personalidad* [Activity, conscience and personality]. Cartago.
- Leontiev, A. N. (2009). La importancia del concepto de actividad objetual para la psicología [The Importance of the concept of objectal activity for psychology]. In L. Quintanar & Y. Solovieva (Eds.), *Las funciones psicológicas en el desarrollo del niño* (pp. 54-63). Trillas.
- Lupyan, G., & Bergen, B. (2016). How language programs the minds. *Topics in Cognitive Science*, 8, 408-424.

- Manzano-Leon, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A Systematic literature review of gamification in education. *Sustainability*, 13(2247). <https://doi.org/10.3390/su13042247>
- Menninger, K. (1992). *Number words and number symbols*. Vanderdoheck & Ruprecht Publishing company. The MIT Press.
- Ministerio de Educación (MINEDU). (2018). *Evaluación PISA 2018* (Diapositivas en PowerPoint) [Assessment PISA 2018]. <http://umc.minedu.gob.pe/resultadospisa2018/>
- Ministerio de Educación (MINEDU). (2021). *Programa curricular de educación primaria* [Primary education curriculum program]. <http://www.minedu.gob.pe/curriculo/pdf/programa-nivel-primaria-ebr.pdf>
- Mouaheb, H., Fahli, A., Moussetad, M., & Eljamali, S. (2012). The Serious game: What educational benefits? *Procedia-Social and Behavioral Sciences*, 46, 5502-5508.
- Pereyra Sánchez, H. (1990). *La Yupana, complemento operacional del quipu* [Yupana, operational complement of quipu]. In C. Mackey, H. Pereyra, C. Radicati, H. Rodríguez & O. Valverde (Eds.), *Quipu y Yupana: colección de escritos* (pp. 235-255). Consejo Nacional de Ciencia, Tecnología e Innovación.
- Prem, D. (2016). *Yupana Inka Decodificando la Matemática Inka* [Yupana Inka decoding the Inka's math – Tawa Pukllay method]. Método Tawa Pukllay.
- Prem, D. (2018a). *Hatun Yupana Qellqa* [The Big book of the Yupana – Decoding the Inka's math]. *Decodificando la Matemática Inka*. Asociación Yupanki.
- Prem, D. (2018b). *Huq, iskay, kimsa...quechua, el idioma computacional de los inkas* [One, two, three... quechua, the computational language of the inkas]. Asociación Yupanki.
- Saldívar, C., Saldívar, A., & Goycochea, D. (2019a). Tawa Pukllay - la aritmética inca de reconocimiento de formas y movimientos operable en paralelo y que no requiere cálculos numéricos mentales [Tawa Pukllay: The inca arithmetic of pattern recognition that works in parallel and doesn't require mental numeric calculations]. En R. Flores, D. García & I. E. Pérez-Vera (Eds.), *Acta Latinoamericana de Matemática Educativa* (pp. 354-363). Comité Latinoamericano de Matemática Educativa.
- Saldívar, C. (2019). P'awaq Yupana–Neoábaco de lógica híbrida [P'awaq Yupana - Hybrid logic Neo abacus]. In *Actas del Congreso Internacional de Ingeniería de Sistemas* (pp. 278-279).
- Saldívar, C. Saldívar, A., & Guzman-Jimenez, R. (2019b). Tawa Pukllay Atipanakuy: The 4 sacred games of the inkas in arhythmic-playful tournament. In *MemoriasCIIE 2019* (pp. 691-696).
- Swacha, J. (2021). State of research on gamification in education: A Bibliometric survey. *Education Sciences*, 11(2), 69. <https://doi.org/10.3390/educsci11020069>
- Tsai, C.-Y., Lin, H.-S., & Liu, S. (2019). The Effect of pedagogical GAME model on students' PISA scientific competencies. *Journal of Computer Assisted Learning*, 36, 359–369.
- Talizina, N. (2009). *La teoría de la actividad aplicada a la enseñanza* [The Theory of activity applied to teaching]. Benemérita Universidad Autónoma de Puebla.
- Urton, G. (2005). *Signos del khipu inka: Código binario* [Signs of the Khipu Inka: Binary code] (Vol. 7). Centro de Estudios Regionales Andinos Bartolomé de Las Casas.
- Vygotski, L. S. (2017). El instrumento y el signo en el desarrollo del niño [The Tool and sign in development of children]. In L. S. Vygotski (Ed.), *Obras Escogidas VI. Herencia científica* (pp. 9-100). Machado Nuevo Aprendizaje.
- Vygotski, L. S. (1995). *Obras Escogidas – III. Problemas del desarrollo de la psique* [Chosen texts – III Problems of psyche development]. Machado Nuevo Aprendizaje.
- Vygotski, L.S. (2010). Aprendizagem e desenvolvimento intelectual na idade escolar [Intellectual learning and development at school age]. In L. S. Vygotski, A. R. Luria & A. N., Leontiev (Eds.), *Linguagem, desenvolvimento e aprendizagem* (pp.103-117). Ícone editora.
- Wassman, J., & Dasen, P. (1994). Yupno number system and counting. *Journal of Cross-Cultural Psychology*, 25, 78-94.
- Wheatley, M. J. (1994). *Leadership and the new science: Learning about organization from an orderly universe*. Berrett-Koehler Publishers.
- Wiese, H. (2007). The Co-evolution of number concepts and counting words. *Lingua*, 117, 758-772.
- Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4), 458–477.

Appendix 1

Test “Yupay Tupay – Sami”

1.- Las matemáticas me hacen sentir... [Mathematics makes me feel...]					
2.- Cuando doy un examen de matemática, me siento... [When I have a math test, I feel...]					
3.- Cuando veo que las cosas que quiero hacer necesitan matemáticas, me siento... [When I see things that I want to do need math, I feel...]					
4.- Conociendo nuevos temas de matemáticas me siento... [Learning new math themes I feel...]					
5.- Cuando estoy estudiando matemáticas me siento... [When I study maths, I feel...]					
6.- Cuando me dan un problema de matemáticas me siento... [When they give me a math problem, I feel...]					
7.- Cuando pienso que las matemáticas tienen muchos más temas por descubrir, me siento... [When I think that maths have many more themes to discover, I feel...]					
8.- Cuando pienso que de grande trabajaré usando las matemáticas, siento... [When I wonder that being older I will work with maths, I feel...]					
9.- Cuando tengo que resolver un problema de matemática solo, me siento... [When I have to solve a math problem alone, I feel...]					
10. Cuando se acerca un examen de matemática, me siento... [When a math test is coming up, I feel...]					

Disfrute [Enjoy]	Motivación [Motivation]	Autoconfianza [Self-confidence]	Valor [Value]	Ansiedad [Anxiety]
---------------------	----------------------------	------------------------------------	------------------	-----------------------

Guest Editorial: Human-centered AI in Education: Augment Human Intelligence with Machine Intelligence

Stephen J. H. Yang^{1*}, Hiroaki Ogata² and Tatsunori Matsui³

¹National Central University, Taiwan // ²Kyoto University, Japan // ³Waseda University, Japan // jhyang@csie.ncu.edu.tw // hiroaki.ogata@gmail.com // matsui-t@waseda.jp

*Corresponding author

ABSTRACT: This special issue focus on underlying research with the use of human-centered AI (Artificial Intelligence), where the new design methods and tools can be leveraged and evaluated, hopes to advance AI research, education, policy, and practice to improve the human condition in education. This special issue intends to advocate an in-depth dialogue between researchers with diverse thoughts, genders, ethnicity, and cultures, as well as across disciplines, leading to a better understanding of human-centered AI. Beneficial interactions between researchers could enhance the adoption of human-centered AI in education. This special issue includes ten papers demonstrating how to augment human intelligence with machine intelligence. The ten papers feature human-centered AI in education, AI in language education, AI in learning analytics, ethical reasoning, AI in the clinical workplace, intelligent education robots, AI risk framework, intelligent course recommendation, education chatbots, and intelligent assessment. Together with the ten papers, we achieve a better understanding of the application of human-centered AI in education.

Keywords: Human-centered AI, AI in education, Humanity, Sustainable education, Future learning

1. Introduction of human-centered AI in education

As we strive to develop AI technology, we also need to reflect appropriately on the impact of social change on education. How do we provide fair and explainable analysis results to gain learners' and teachers' trust? How do we guide learners and teachers to meet challenges from both technical and organizational aspects? How do we consider technological development with social value and work towards sustainable education and future learning?

The advance of AI in decision-making, prediction, knowledge extraction, and logic reasoning has been making a broader impact on society, the economy, and the environment (Luan et al., 2020; Yang et al., 2021). AI has the potential to educate, train, and augment human productivity, making them better at their tasks and activities. AI can also make the better quality of an individual's work, resulting in better learning and teaching. Human-centered AI can be interpreted from two perspectives (Yang, 2021; Yang et al., 2021), one is AI under human control as addressed by Shneiderman (2020), and the other is AI concerning the human condition defined by Stanford HAI (2022). AI under human control is to leverage the collaboration between human control and AI automation to empower human productivity with high reliability, safety, and trust. AI concerning the human condition is that AI algorithms taking humanity as the primary consideration, require explainable and interpretable computation and judgment process, and continuously adjust AI algorithms through human context and societal phenomena to augment human intelligence with machine intelligence, thereby enhancing the welfare of human kinds.

The shifting of AI research trends has brought new applications of AI in education. One example of transfer intelligence is the generation and adoption of new deep learning algorithms with pre-trained knowledge datasets in natural language processing, such as BERT with 340M dataset mentioned in Devlin et al. (2018), GPT-3 with 175B dataset in Brown et al. (2020), and Megatron-Turing NLG with 530B dataset in Smith et al. (2022). They apply pre-trained knowledge to fine-tune domains and be more effective than the previous generation of deep learning and traditional machine learning algorithms. These new algorithms can achieve performance that is closer to humans. In addition, the promise of precision education commits to applying AI research to intelligent tutoring for precise adaption and personalization, precise profiling, diagnosis, prediction, treatment, and prevention for smart assessment and evaluation as mentioned in Yang et al. (2022).

In addition, potential ethical issues are involved as AI requires a large amount of learner data, sometimes sensitive information for model training. The data collection process must obtain the consent of the students and teachers in the first place, and the management and storage of data must meet the requirements of data security and the protection of personal privacy to make AI educational systems theoretically and educationally sound.

2. Human-centered AI toward sustainable education

Sustainable education is a quality education considering humanity. The challenge for sustainable education is incorporating cultural and social changes into the design from the outset, including educating all stakeholders and providing the appropriate training. To develop the most helpful strategies for stakeholders from different perspectives, such as the content, methods, tools, and platforms for education and training.

The theme of human-centered AI toward sustainable education includes ethical issues of fairness and equality, explainable and trustworthiness; social issues of diversity and inclusion, resilience and robustness; governance issues of accountability, data safety, adaptation, and accessibility.

AI could be misused because of biases in data and algorithms. Analytic algorithms trained on regular articles will learn and reproduce the societal biases against women and minorities, which are embedded in languages and culture. Word embedding is an example, it is a popular technique in natural language processing has been found to exacerbate existing gender and racial stereotypes. Fairness and equality means that the analysis technology must produce unbiased and fair results. The analysis process should not include discrimination and unfair analysis results against race, religion, gender, and physical disability. We can avoid bias of data/algorithms by understanding cultural and social impact on education, and by designing bias detection and prevention algorithms in the pre-processing, in-processing, and post-processing of AI model training.

Human-centered AI must have sufficient interpretability, and the current algorithms are inadequate in this regard. Explainability provides a certain degree of transparency and explanation in the decision-making process. Explain the process of reaching conclusions and adjusting the transparency of data and algorithms according to the differences of stakeholders. Trust comes with accuracy, transparency, explainable, and fairness. This decision-making process requires complete explanations to gain trust and avoid unnecessary negative consequences. Therefore, explainable and trustworthy AI is necessary to enable explanation and comprehension so humans can understand how AI makes decisions. Researchers are working on explainable algorithms, hoping they can explain the reasons for each decision to increase the trustworthiness when making complex decisions.

Inclusion is based on diversity, equity, and belonging. Inclusive education breaks down systemic barriers to inclusion. It fosters a culture where every learner knows their belonging, feels empowered to bring their whole self to learning, and is inspired to learn. When we face sudden and dramatic changes in our living and educational environment, resilience reflects how we can recover from natural disasters or disease pandemics like COVID-19. Resilience education includes the technical robustness and safety of networks and devices, accessibility of teachers, and adaptation and accessibility of content, tools, and platforms.

With human-centered AI considering fairness, equality, inclusion, diversity, explainability, trustworthiness, and resilience, we can work together toward sustainable education.

3. Contribution of papers to this special issue

Ten papers have been included in this special issue. They address how to achieve the goal of human-centered AI in education and why their proposed system and method are better while considering humanity. Papers in this special issue inspire future studies of human-centered AI and conclude the finding in their study based on analyzing data collected from experiments or a systematic review. The ten papers feature human-centered AI in education, AI in language education, AI in learning analytics, ethical reasoning, AI in the clinical workplace, Intelligent education robots, risk framework, intelligent course recommendation, education chatbots, and intelligent assessment, together to achieve a better understanding of the application of human-centered AI in education. The following is the list of papers' titles and authors.

Title: Unpacking the “Black Box” of AI in Education,

Authors: Nabeel Gillani, Rebecca Eynon, Catherine Chiabaut, and Kelsey Finkel

Title: Trends, Research Issues, and Applications of Artificial Intelligence in Language Education

Authors: Xinyi Huang, Di Zou, Gary Cheng, Xieling Chen, and Haoran Xie

Title: A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior

Authors: Fuzheng Zhao, Gi-Zen Liu, Juan Zhou, and Chengjiu Yin

Title: A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning

Authors: Alwyn Vwen Yen Lee, Andrés Carlos Luco, and Seng Chee Tan

Title: Feasibility and Accessibility of Human-centered AI-based Simulation System for Improving the Occupational Safety of Clinical Workplace

Authors: Pin-Hsuan Wang, Anna YuQing Huang, Yen-Hsun Huang, Ying-Ying Yang, Jiing-Feng Lirng, Tzu-Hao Li, Ming-Chih Hou, Chen-Huan Chen, Albert ChihChieh Yang, Chi-Hung Lin, and Wayne Huey-Herng Sheu

Title: Artificial Intelligent Robots for Precision Education: A Topic Modeling-Based Bibliometric Analysis

Authors: Xieling Chen, Gary Cheng, Di Zou, Baichang Zhong, and Haoran Xie

Title: A Risk Framework for Human-centered Artificial Intelligence in Education: Based on Literature Review and Delphi–AHP Method

Authors: Shijin Li and Xiaoqing Gu

Title: AI, Please Help Me Choose a Course: Building a Personalized Hybrid Course Recommendation System to Assist Students in Choosing Courses Adaptively

Authors: Hui-Tzu Chang, Chia-Yu Lin, Wei-Bin Jheng, Shih-Hsu Chen, Hsien-Hua Wu, Fang-Ching Tseng, and Li-Chun Wang

Title: Effects of Incorporating an Expert Decision-making Mechanism into Chatbots on Students' Achievement, Enjoyment, and Anxiety

Authors: Ting-Chia Hsu, Hsiu-Ling Huang, Gwo-Jen Hwang, and Mu-sheng Chen

Title: Application of Artificial Intelligence Techniques in Analysis and Assessment of Digital Competence in University Courses

Authors: Tzu-Chi Yang

4. Conclusion and future research

Modern learning technologies need to be more accurate and intelligent to help students formulate practical learning guidance and intervention. We envision that future learning will closely rely on some fundamental learning technologies, such as smart learning analytics, precision education, and human-centered AI.

Smart learning analytics is a research field with the optimal goal of improving learning and teaching by building better pedagogies, empowering active learning, targeting at-risk students, providing intervention, and assessing student success. The goal is to improve teachers' teaching quality and students' learning outcomes. Research on smart learning analytics is needed to improve the quality of teaching (i.e., teachers must identify and address topics of concern to students, such as inadequate feedback from learning environments), identify which students are struggling with a particular topic, and understand how their content has been used and how effective it is. Smart learning analytics enables teachers to continually enhance educational content to be tailored to students' level of understanding as they progress and monitor student's performance so that teachers can adapt their teaching. Smart learning analytics enables students to take control of their learning, know how they are performing compared with peers, and complete assessments to keep up with the learning progress of their peer group and helps teachers identify gaps in students' prerequisite knowledge and key study skills.

Precision education is to discover students' differences and individual characteristics and guide students to conduct individualized learning accordingly. Teachers can make preventive adjustments to students' critical behaviors. Based on the student's learning status, learning ability, and other relevant individual characteristics, promptly carry out individualized remedial activities. Precision education is the best opportunity to achieve individualization and turn personalized learning from one-size-fits-all to one-of-a-kind. Precision education is to identify at-risk students as early as possible and provide them with timely intervention through diagnosis, prediction, treatment, and prevention. To be more specific, precision education's process diagnoses students' engagement, learning patterns, and behavior. Making predictions concerning students' learning performance and improving predictive models, followed by treatment with learning strategy and activities through timely intervention and prevention. Through precision education, teachers can understand students' learning situations by diagnostic system, extract data and establish a learning prediction model, then design adaptive learning

activities for different types of students with one-of-a-kind treatment and prevention. The challenge ahead is how to accurately establish student and data models so that teachers can better understand students' differences to generate individualization. Establishing a student model involves whether the acquisition of student data is ethical, whether the evaluation of learning outcomes is objective and fair, and whether the student model's establishment is open and transparent.

Future learning is a process of unlearn and relearn to foster student-centered learning. Teachers need to reimagine the future world, unlearn the lecture-oriented teaching method and relearn the human-centered technology to guide students to reimagine the future world. Teachers must also relearn modern learning technology and change from teaching to guiding students to conduct individualized, self-regulated, autonomous, and seamless learning. There are no magic pills in education, like rehabilitation in medical; good teaching needs well-designed strategies and practices. Strategies are diversified, and the value of teachers, like the value of coaches, lies in knowing how to apply learning activities and teaching methods wisely. With the research of smart learning analytics, precision education, and human-centered AI, we envision a pathway toward sustainable education in the future.

Acknowledgment

The guest editors would like to express our sincere thankfulness to Prof. Nian-Shing Chen of National Taiwan Normal University, Taiwan, and Prof. Maiga Chang of Athabasca University, Canada, for their inspiring advice on this special issue. Given the high volume of submissions, this special issue runs with an acceptance rate lower than 25%, and only a few papers that are most related to "Human-centered AI in education" are included. We truly appreciate all the authors and reviewers for their valuable contribution and support towards the success of this special issue and *Educational Technology & Society*.

References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint ArXiv:1810.04805.
- Luan, H., Geczy, P., Lai, H., Gober, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.C. (2020). Challenges and future directions of Big data and artificial intelligence in education. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2020.580820>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504
- Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R., Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (2022). *Using Deepspeed and Megatron to train Megatron-Turing NLG 530b, A Large-scale generative language model*. PsyArXiv. <https://doi.org/10.48550/arXiv.2201.11990>
- Stanford HAI. (2022). *Stanford institute for human-centered artificial intelligence*. <https://hai.stanford.edu/>
- Yang, C. C. Y., Chen, I.Y., Ogata, H. (2022). Toward precision education: Educational data mining and learning analytics for identifying students' learning patterns with ebook systems. *Educational Technology & Society*, 24(1), 152-163.
- Yang, S. J. H. (2021). Guest editorial: Precision education – A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>

Unpacking the “Black Box” of AI in Education

Nabeel Gillani^{1*}, Rebecca Eynon², Catherine Chiabaut³ and Kelsey Finkel³

¹Massachusetts Institute of Technology, USA // ²University of Oxford, UK // ³The Robertson Foundation, USA // ngillani@mit.edu // rebecca.eynon@oii.ox.ac.uk // catherine.chiabaut@robertson.org

// kelsey.finkel@robertson.org

*Corresponding author

ABSTRACT: Recent advances in Artificial Intelligence (AI) have sparked renewed interest in its potential to improve education. However, AI is a loose umbrella term that refers to a collection of methods, capabilities, and limitations—many of which are often not explicitly articulated by researchers, education technology companies, or other AI developers. In this paper, we seek to clarify what “AI” is and the potential it holds to both advance and hamper educational opportunities that may improve the human condition. We offer a basic introduction to different methods and philosophies underpinning AI, discuss recent advances, explore applications to education, and highlight key limitations and risks. We conclude with a set of questions that educationalists may ask as they encounter AI in their research and practice. Our hope is to make often jargon-laden terms and concepts accessible, so that all are equipped to understand, interrogate, and ultimately shape the development of human-centered AI in education.

Keywords: K-12 education, Artificial intelligence in education, Educational data mining, Learning analytics, Natural language processing

1. Introduction

Rapid advances in artificial intelligence (AI) over the past several years have raised new questions about the role that machines might play in both promoting and impeding humanity. The field of education has been no different. Emerging AI capabilities are enabling machines to fuse and make sense of larger, more diverse datasets in increasingly efficient ways. While these affordances of scale, diversity, and efficiency might help generate insights and guide actions to improve educational opportunities and outcomes, they also come with several technical limitations and related practical risks—like failures to generalize and identify causal relationships—that threaten to perpetuate unfair or harmful applications. Thus, and rightfully so, the re-emergence of AI has sparked new debates about the political, pedagogic, and practical implications of its application in educational contexts (Shum & Luckin, 2019). These debates are critical, especially if we wish for machines to be able to better-serve the human actors—teachers, learners, administrators, and others in education—who may benefit from their emerging capabilities.

Engaging productively in these debates, however, requires one to understand some of the methodological paradigms and practices specific to Artificial Intelligence in Education (AIED). However, researchers and practitioners not trained in computer science or engineering may find the rapidly advancing field of AI inaccessible. In this article, we try to address this gap, providing an overview of the meanings, methods, and limitations of AI as a re-emerging field and how these intersect with AI’s applications in education. In doing so, we hope to build on previous introductions to this topic (e.g., Luckin, 2018; Holmes et al., 2019) and critical works that connect data models with ethical and social debates (Perrotta & Williamson, 2018; Perrotta & Selwyn, 2020). By opening up the “Black Box” of AI for those outside of the field, we hope to further human-centered AI in education by empowering all stakeholders, regardless of disciplinary background, to contribute to the development of AI that recognizes and champions human capabilities (Li & Etchemendy, 2018; Yang et al., 2021).

2. Defining “AI”

As Artificial Intelligence evolves, the term “AI” has acquired mystical and rhetorical qualities (Eynon & Young, 2020). Some recent advances are impressive: we now have machines that can discover new drug formulas (Popova et al., 2018), predict elusive protein structures (AlphaFold Team, 2020), generate full-length written stories (Brown et al., 2020), and beat world-class performers in games like Starcraft, Go, and Chess (AlphaStar Team, 2019; Silver et al., 2018). Still, demystifying AI is an important first step towards understanding its inner workings and applications. While the capabilities and performance of today’s AI systems are unprecedented, many of the core algorithms that govern how they work are rooted in methods dating back to the early 20th century (Tuomi, 2018). Furthermore, while current incarnations of AI have achieved unprecedented degrees of

sophistication, the “I” of AI systems remains quite rudimentary—as evidenced by how poorly these systems often perform on tasks that humans find intuitive. Such technical limitations entail important risks and ethical considerations which have significant bearings on the application of AI to the field of education. Before delving into these risks, we expand on two schools of AI that are frequently used in education—machine learning and rule-based AI—and outline some of their common applications.

2.1. Machine learning-based AI

2.1.1. Machine learning paradigms: supervised, unsupervised, and reinforcement learning

Machine learning algorithms are designed to mine large datasets to uncover—or “learn”—latent rules and patterns that may help inform some future decision. For example, imagine a large school system has asked a research team to develop a tool that accurately predicts what a student’s GPA will be at the end of a given school year. “Supervised learning” is one approach to machine learning that could help them tackle this problem. With supervised learning, machines are provided a historical dataset of inputs, or features (e.g., student-level characteristics like demographic data, attendance records, test scores), along with a target output, or attribute (e.g., GPA). A model is then applied to the dataset to learn how these features map to the target attribute by testing out different hypotheses about the relationship to or path from student-level characteristics to GPA. The labels (historical GPAs) of the data in the set help “supervise” the model by indicating how far off its predictions are from the observed or existing (i.e., ground-truth) values. This occurs iteratively for each data point, eventually “training” the model by updating the weights it attaches to the inputs or other variables it uses to make predictions. These weights are often the quantities “learned” by the machine (hence, the term “machine learning”). Linear regression offers a classic approach to supervised machine learning. In fact, many modern approaches using neural networks (described in more detail below), while often described in quasi-mystical terms in press articles, operate in fundamentally similar ways and seek to achieve similar outcomes as linear regression. The scenario in the Supplementary Materials offers additional details about these similarities and differences.

In contrast to supervised learning, “unsupervised learning” is a process by which a machine performs statistical pattern recognition without access to ground-truth labels for the desired output. A common application of unsupervised learning is clustering. Say a school system asked a research team to develop a “typology” of students based on their different characteristics, to help design and target student supports. They could use a standard clustering algorithm (e.g., the popular “k-means” algorithm proposed by Hartigan and Wong, 1979) to learn a grouping of students that differentiates them from other (also automatically inferred) groups. Our resultant groups—or clusters—may comprise students who perform similarly; who take similar classes; live in similar parts of the city; or have some other set of related characteristics.

A third paradigm of machine learning is “reinforcement learning,” which has recently been used, among other applications, to develop powerful gameplay systems (e.g., Mnih et al., 2015; Silver et al., 2018). In education, some researchers have started to explore applications of reinforcement learning to intelligent tutoring systems (Reddy et al., 2017). At its core, a reinforcement learning algorithm accepts as an input the state of the world (e.g., the questions a student has answered correctly or incorrectly in an intelligent tutoring environment of a game) and uses this to decide upon some action (e.g., which question to ask the student next). The action—either immediately or over the course of time—eventually contributes to some outcome (e.g., mastering a concept). The value of this outcome is then used to assign positive or negative rewards to the algorithm to encourage or discourage similar actions when faced with similar states of the world in the future. Reinforcement learning algorithms have been around for several decades (Kaelbling et al., 1996), but have resurged over the past few years with large quantities of training data and computational resources more readily available.

2.1.2. Machine learning philosophies: frequentist and Bayesian

The paradigms above reflect a “frequentist” philosophy of machine learning: inferences (like predictions, cluster assignments, and other insights that inform decisions) are made largely based on the frequencies of patterns revealed in the training data (Bayyari & Berger, 2004). By contrast, “Bayesian” machine learning models explicitly incorporate pre-existing beliefs (“priors”) alongside the patterns revealed by training data to produce some posterior “belief” or inference about the world (Bayyari & Berger, 2004).

Say, for example, a biased coin is tossed 100 times and yields heads on 30 instances. Two friends make a bet that they can infer the true bias of the coin and predict the next 100 tosses. One trains a frequentist machine learning model on the observed coin tosses, which simply factors in the observed data. The model infers the bias as equalling a 30% likelihood of landing on heads. The other friend, however, devises and trains a Bayesian model: in addition to factoring in the observed number of heads, she also factors in a prior belief drawn from most normal coin tossing activities: that there is distribution of possible chances that the coin will land on heads (centered around what we usually expect from coins, 50%). On the next 100 tosses, both observe 40 heads. In this instance, factoring in prior beliefs into the model—instead of simply trusting the observed data—produced an inference of the coin’s bias as falling between 30% and 50%, which was more accurate than trusting only the data from the initial set of coin tosses.

Since they rely on both observed data *and* prior beliefs, Bayesian methods can sometimes help overcome sparsity in datasets—like our limited number of coin tosses—in order to make more accurate predictions. In other cases, such prior beliefs may themselves be biased and therefore make models less accurate than if they were trained only on observations. Whether a Bayesian or frequentist model is more appropriate to use depends on the nature of the problem at hand. Interestingly, many believe that the rich structure of Bayesian models reflects aspects of human cognition (Tenenbaum et al., 2011), making them “truer-to-nature” AI. However, many methods for conducting Bayesian posterior inference do not scale well to large datasets, making them difficult to deploy in several real-world settings. In practice, many approaches to machine learning can be implemented from either a Bayesian or Frequentist point of view.

2.1.3. The rise of deep learning

Deep learning—a popular approach to machine learning—has become the dominant school of AI in recent years owing largely to a resurgent interest in neural networks. Neural networks take inspiration from connectionist philosophies of cognitive science (Elman et al., 1996) and generally operate by learning (possibly nonlinear) relationships between several input variables in order to produce predictions as accurately as possible (see the machine learning scenario in the Supplementary Materials for more details). They are the core, modular building blocks that make deep learning systems “deep”: combining smaller neural networks together to form larger ones by feeding the outputs of one as inputs to another can enable the discovery of more complex and granular relationships between these inputs and outputs (LeCun et al., 2015). Neural networks can manifest through a number of different algorithmic architectures, e.g., Recurrent Neural Networks (RNNs (Goodfellow et al., 2016)), Convolutional Neural Networks (CNNs (Goodfellow et al., 2016)), and Transformers (Vaswani et al., 2017)—which underpin recent advances in natural language processing like the popular BERT model (Devlin et al., 2019). Each of these architectures differ in how they process and transform inputs into outputs. Furthermore, RNNs and Transformers are generally better-suited for tasks that involve time-series data, whereas CNNs are often applied to image processing problems. Still, while their precise structures and implementations may differ, many of these architectures are trained, evaluated, and eventually used in similar ways.

Deep learning has been driven by advances across three major areas over the past several years: data, algorithms, and hardware. Large, easy-to-access datasets have enabled, for example, the recent “GPT-3” language model—which is trained on over 570 gigabytes of text found across the open internet (Brown et al., 2020). The model is simply trained to predict the next word in a corpus of text given some sequence of preceding words. The result is a powerful system that can generate entire believable stories—an exciting possibility, but also of particular concern in our current era of misinformation (OpenAI Team, 2019).

In cases where large datasets are not available for a specific task, algorithmic advances like “transfer learning” can help (Pan & Yang, 2009). Transfer learning enables a model to “pre-train” itself—i.e., initialize its parameters—using the outputs of a training process conducted for a separate but related task for which enough data *is* available. The model can then “fine-tune” on—or adapt itself to—a smaller dataset that more closely represents the task at hand. For example, early warning systems to detect students likely to drop out may be developed for districts that lack a breadth or depth of historical data by “borrowing” the predictive capacities of models pre-trained on data from larger school settings as a starting point (Coleman et al., 2019). Pre-training, however, may also contribute to the amplification and propagation of biases across models.

Finally, recent hardware accelerations like Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs, Hennessy & Patterson, 2019) are enabling more time-efficient computation, yet their energy demands and associated costs have raised concerns about their potential environmental impacts (García-Martín et al., 2019) and contribution to widening divides between the AI capabilities of large companies and smaller research groups (Hao, 2019).

2.2. Rule-based AI

Machine learning systems can be powerful, particularly for problems where the “rules” needed to produce certain outcomes (e.g., the weights to be applied to students’ characteristics in order to produce GPA predictions) are not known and hence must be inferred from data. However, there are also problems for which the rules *are* known, but applying these rules can be cumbersome or time-consuming. For these types of problems, “rule-based” approaches to AI—in which computers manipulate data based on a set of pre-defined logical propositions, instead of ones inferred from patterns in the data—are often used.

One such problem in education is school bus routing. Large school districts often have fleets of school buses that must be scheduled and routed to different stops in order to ensure students get to school on time and safely (Bertsimas et al., 2019). In this problem, the rules an AI would consider might include: different carrying capacities for each bus; times by which certain groups of students must get to school; or specific roads buses can and cannot take. A social planner implementing this algorithm may seek to optimize for multiple objective functions: for example, minimizing costs and travel times and/or maximizing the diversity of the student body that travels together on any given bus.

The most naive rule-based AI algorithm would use a brute force approach to solve this problem, evaluating every possible combination of bus, student, and route assignments and selecting the one that yields the most optimal response vis-a-vis our multiple objectives. For many large real-world problems, however, this approach is infeasible and could quite literally take hundreds of years (or longer) to compute (Cook, 2012). To this end, rule-based AI algorithms often use sophisticated solution strategies to prune down a large set of possible combinations to a feasible subset that is much easier and more efficient to search through (e.g., Van Hentenryck & Michel, 2009). Unlike machine learning systems, rule-based models will not necessarily make more accurate decisions with a larger scale or diversity of data. In fact, scale and diversity of data can pose challenges to rule-based AI algorithms because they increase the size and complexity of the problem at hand. This said, these challenges will likely be alleviated by the increased algorithmic and hardware efficiencies afforded by the current wave of AI described above.

While their underlying mechanisms might differ, rule-based AI need not be completely distinct from machine learning. For example, we may have historical data on bus routes and road conditions (e.g., traffic patterns) which we can use to predict travel times. We can then leverage these predicted travel times as inputs into our objective function during the optimization process.

3. Applications of AI in education

Despite recent interest in applications of AI and education, the two fields have intersected for some time (e.g., Aleven & Koedinger, 2002)—which has long raised important philosophical and ethical questions. This next section provides an overview of recent applications of AI in education and highlights some of their limitations and broader implications. These examples, far from exhaustive, have been selected in order to highlight the ways in which the scale and diversity of available data—along with improvements in computational efficiency—have created new opportunities for using AI to potentially improve the human condition through educational applications. For a more in-depth review of how AI and other data mining techniques *can* be applied to education-related problems, we refer readers to several existing review papers (e.g., Romero & Ventura, 2010; Koedinger et al., 2015; Fischer et al., 2020).

3.1. Intelligent tutoring systems

Intelligent tutoring systems (ITS) are a popular application of AI in education. ITS are tools that seek to adapt to students’ existing knowledge and skills, or learning states, to help them build skills in more personalized ways. The “I” in ITS often has different definitions for different tools. For example, some ITS are machine learning-based systems that seek to develop (sometimes Bayesian) learner models trained to maximize the likelihood of a student answering a provided question correctly, conditional on their history of responses (Ritter, 2007). In other cases, developers might simply train a system to predict the likelihood of “correctness” as accurately as possible (e.g., using deep reinforcement learning a la Reddy et al., 2017). These systems then provide students with problems that are most likely to be at their “learning edge”—i.e., the problems they haven’t yet answered that they are most likely to answer correctly, given their prior history of answers. These machine learning systems have the capacity to make more accurate predictions of a student’s learning edge as they draw on larger and more

disparate historical sources of student performance and behavior—many of which are becoming more ubiquitous through computer-aided tutoring and assessment platforms. Other ITS, like (Kelly et al., 2013), pre-define simple rules—like correctly answering three similar questions in a row—to determine if and when a student has mastered some concept.

Experimental evidence has largely shown ITS to be effective in increasing students' grades and test scores (J-PAL Evidence Review, 2019). Of course, grades and test scores offer only one (limited) view into student learning. Crucially, much of the existing efficacy research on ITS has not specifically analysed which underlying AI methods make them more or less effective. As such, it is unclear to what extent machine learning vs rule-based systems are responsible for helping students improve their outcomes. As machine learning technologies continue to offer new opportunities for personalizing instruction, it will be important to identify the precise elements of these systems that offer the greatest promise for enhancing student learning. There is also a need to better understand the contexts in which these ITS systems can be meaningfully deployed as a resource for teachers and students in ways that do not inadvertently narrow the aims and purposes of Education (Biesta, 2015).

3.2. Assessment and feedback

Proponents of AI, particularly machine-learning based systems that seek to infer students' knowledge states from the growing scale and diversity of data available on digital learning platforms like Khan Academy, argue these systems have the capacity to obviate the need for explicit formative and summative assessments, by seeking to infer students' knowledge states from the growing scale and diversity of data available on such digital learning platforms (Piech et al., 2015) and other systems instrumented for “learning analytics” (Gašević et al., 2015). After all, if it is possible to know what a student knows based on how they answer questions in an ITS, why administer an assessment at all? This line of reasoning, of course, does not consider the positive effects exam preparation and studying can have on learning (Karpicke & Roediger III, 2008).

Automated assessment of writing submissions is a popular, albeit complex, example of how machine learning might support assessment. To date, most research has focused on training machine learning models to assess foundational attributes of writing—for example, spelling, vocabulary, and grammar. Other systems have used machine learning to train models that are able to replicate human scores for a given essay (Dong et al., 2017). Growing as a writer, however, requires much more than feedback on the mechanics of writing or collapsing a rich composition down to a single grade. To this end, (Fiacco et al., 2019) recently designed a neural network-based machine learning system to identify which rhetorical structures were present in sentences contained within a corpus of research study articles: for example, which sentences sought to describe the study, provide context on the study's methods, or frame new knowledge.

Despite the advancing capabilities of these systems, however, some concerns remain. For instance, it would be important to train these AI on a diverse set of linguistic data to fuel their accuracy and minimize bias. More work also needs to be done to understand how they might inadvertently negatively impact writing development and written work in the same ways as plagiarism detection software has (Ross & Macleod, 2020), and more generally, how student surveillance via constant data collection may impact students (Eynon, 2013). Thus, although assessment and feedback is a core focus of AIED, the most appropriate ways to deploy AI for particular activities and in specific contexts remains an area of debate.

3.3. Coaching and counselling

The role of coaches and counsellors in schools are multifaceted, time-intensive, and costly. Researchers have therefore started to explore how some of their tasks can be automated. For example, several studies have demonstrated how text-message reminders can help facilitate specific outcomes normally under counsellors' purview: e.g., ensuring that graduated high school seniors take the steps needed to matriculate at college in the fall (Castleman & Page, 2015) and keeping parents updated about their children's academics (Bergman & Chen, 2019).

Recent efforts have also leveraged AI to enable a richer set of interactions between students and “counselors.” A recent study (Page & Gehlbach, 2017) deployed an AI chatbot to answer questions about forms students would need to fill out before starting college at Georgia State University (GSU). The authors indicate that the chatbot was trained using deep reinforcement learning—the same technology that has enabled state-of-the-art advances

in automated gameplay (AlphaStar Team, 2019)—though the exact methods for training and evaluating these models in the context of the chatbot are unclear. The researchers found that the AI-powered system was comparable in enhancing college enrolment rates to prior studies that primarily involved human counselors. As more dialogue agent systems are deployed across campuses, the scale and breadth of available linguistic corpora for training models with smarter response strategies are likely to grow. Nevertheless, a number of open research questions persist—particularly concerning how well these systems can serve a diverse student body in answering complex educational questions.

3.4. (Large) school systems-level processes

At the school systems-level, AI is being used to achieve several objectives, including the equitable implementation of school choice. Over the past two decades, a strand of economics research has focused on developing rule-based AI algorithms for districts that offer families choices on where to send their children to school. These algorithms have been designed to be “strategy-proof,” matching students to schools in ways that do not enable families to “game the system” by mis-stating preferences in order to exploit loopholes that would increase their likelihood of receiving a spot at one of their top choice schools (Pathak & Sönmez, 2008). This is particularly important for those parents who do not have the resources, social capital, or knowledge necessary to “game the system.” Of course, “strategy-proofness” only helps further equity to the extent that other parts of the system are also equitable (Goldstein, 2019).

AI has also been used to help with a range of planning and forecasting tasks, particularly in larger school-systems or by those working across large systems of schools. Working with Boston Public Schools, researchers built a machine learning model that forecasts changes in demand for schools in response to certain school choice policy changes (Pathak & Shi, 2015). As more data accrues across the diverse spectrum of families in these systems, such models have the potential to become more accurate—and perhaps also shed more light on the preferences of families who belong to traditionally underrepresented segments of the population. School districts have also turned to rule-based AI systems to help achieve greater logistical efficiency—for example, by producing “optimal” bus routes as discussed earlier in this paper—and to save money (Bertsimas et al., 2019). Yet such systems have been met with mixed reception from some of the families they ultimately impact (Scharfenberg, 2018). Additionally, to improve teacher placement in schools, Teach for America (TFA) designed and tested a matching algorithm similar to the school choice matching algorithm described above, to factor in both teacher and school preferences; TFA subsequently saw a slight positive effect on students’ academic outcomes (Davis, 2017). With continued increases in computational efficiency, these rule-based systems promise to be able to operate on larger, more complex problems concerning more students, teachers, and other stakeholders in the years ahead. Yet these need to be developed with an awareness of concerns about the use of such market-driven principles to develop an equitable education system (e.g., Ball, 2017; Biesta, 2015).

3.5. Predicting outcomes

Machine learning systems have garnered significant attention for their ability to “predict the future”—often in the form of “early warning systems.” These systems, often using different forms of regression, mine large troves of historical student data to predict which students are most at risk of failing an exam, dropping out of high school or college, etc. (Faria et al., 2017). Experimental evidence has suggested that deploying these systems can help reduce chronic absenteeism and course failure (Faria et al., 2017). While early warning systems do not always require machine learning—e.g., a simple rule-based system could trigger a warning if a student’s GPA falls below a certain level—machine learning-based systems have the potential to identify and exploit patterns of which school leaders may not be aware. These systems can also pool data across disparate contexts to improve individual predictions. For example, small school districts might face a “cold start” problem: they simply do not have enough historical data to train an accurate machine learning model—requiring them to “borrow” data from other school districts to improve accuracy (e.g., Coleman et al., 2019). Increasing scale and diversity of data may enable such applications of transfer learning, and more generally, extend the possible applications of machine learning to educational settings that have previously been left out.

Unfortunately, these warning systems can have several drawbacks. Being able to predict how well a student is going to do in a particular class might help encourage students to take more advanced classes (Bergman et al., 2021)—but it could also lead to tracking, which might limit a student’s desire and ability to explore new topics, particularly in college and university. School leaders may also struggle to calibrate interventions based on the outputs of a model. If a model indicates the probability that any given student drops out of high school, at what

point should an intervention be triggered—when there is a 20% chance of a student dropping out? 51%? 90%? Even if a school leader feels equipped to intervene after analyzing the data, there is a fundamental question about the obligation to act (Eynon, 2013; Prinsloo et al., 2017; Hakimi et al., 2021): which students should receive support? And what if the model has a high false-negative rate—meaning there could be many students who actually need intervention but weren't flagged by the model as such? These are difficult questions and, at present, there are no standardized responses; school systems approach these questions differently depending on their own knowledge and needs.

4. Limitations and risks of modern machine learning systems

Readers from varied sub-fields of education, learning sciences and data science will bring different critical lenses to the areas and applications previously discussed. Here, we will draw from (Lake et al., 2016) and other researchers to discuss several technical limitations of modern machine learning systems and some risks that arise from them. We will also look at the key gaps that still exist between what many believe AI can do in 2021, what it can actually do (and not do), and how these limitations have important implications for education.

4.1. Limitations of modern machine learning systems

4.1.1. Transparency and interpretability

Neural network approaches to machine learning are powerful, but their inner workings are usually not transparent, making them difficult to interpret. One implication of this is that it may not be clear which inputs were responsible for driving decisions. For example, in the case of early warning systems, a school leader might be informed of the likelihood of any given student failing a course, but not which characteristics of the student are most associated with this prediction. The school leader might obviate this problem by opting for a more interpretable, non-deep learning-based model, but this may require sacrificing some degree of predictive accuracy. These are not always salient tradeoffs, but when they arise, it is often unclear how they should be made. Fortunately, model interpretability is an active area of deep learning research with several recent advances (e.g., Sundarajan et al., 2017; Kim et al., 2018). These advances are critical for equity and inclusion in education, as they open the door to enabling a wider range of stakeholders—including parents and students who may be affected by such algorithms—to understand, interrogate, and ultimately improve their applications (although see Ananny & Crawford, 2018; Tsai et al., 2019 for discussions of questions of the burden such moves could place on individuals).

A more fundamental issue with machine-learning based systems, even those that do not leverage deep neural networks, is causal attribution. Machine learning models are designed to identify and exploit correlations (not necessarily causal relationships) between variables in order to make predictions. For example, a school leader's early warning system might highlight poverty status, prior grade history, and disciplinary actions as student-level factors associated with a higher likelihood of course failure, without explaining the underlying *causes* of failure. Misunderstanding underlying causes may lead to faulty or incomplete interventions, and ultimately, a perpetuation (or exacerbation) of the underlying educational challenges educationalists are seeking to address. Advances in machine learning methods for causal analysis (e.g., Johansson et al., 2016) are attempting to help separate out correlation from causation. However, grasping a rich understanding of causal processes in settings as complex as education usually requires much more than technical solutions.

4.1.2. Abstract reasoning and learning how to learn

Humans are very good at two things that AI-powered machines are not: abstract reasoning and learning how to learn. For example, while machines can learn to play a variety of games better than champion-calibre players, they require training on simulations of hundreds of thousands or millions of games to learn how to do so. Humans, by contrast, often learn gameplay simply by watching someone else play for a few minutes (Lake et al., 2016). This is partly because we are remarkably adept at abstract reasoning: ascertaining the fundamental rules of a particular task to generalize and apply these rules to other similar but distinct endeavors. Teachers do this all the time: unlike most intelligent tutoring systems, they do not need to observe a large number of question responses from a student in order to identify and begin addressing key conceptual gaps.

An important type of abstract reasoning is learning how to learn. Throughout our lives, we have likely played several games, and these experiences have made it easier for us to learn the rules and dynamics of new games. Such “meta learning” is a popular area of machine learning research. At present, however, the complex reasoning done by humans is broken down into discrete processes for the machines, including teaching a machine “where to focus” in the space of input data (Xu et al., 2015) or how to automatically update different parts of its own architecture (Andrychowicz et al., 2016; Zoph & Le, 2016) in order to make better predictions. Perhaps unsurprisingly, this machine “meta learning” generally lacks the higher-order thinking, reflection, and planning woven throughout human meta-learning. Without this ability to learn how to learn, we must be skeptical of how well AI can support students, understanding the complex in- and out-of-school factors that impact learning. AI, for example, may be able to suggest problems to students to work on, but will be limited in identifying why students continue to get certain types of problems right or wrong—especially if those factors transcend cognitive, skill-based challenges and extend to the home environment or other social forces affecting the child.

4.2. Risks that stem from machine learning’s limitations

4.2.1. Failures in generalizing

Because machine learning models often fail to develop a deep, intuitive understanding of the task they are built to perform, they can subsequently fail to generalize to new settings than what they were trained for (Murphy, 2012). This sometimes leads them to “catastrophically forget” how to perform tasks (Kirkpatrick et al., 2017), or become brittle in the face of “adversarial” inputs. Adversarial inputs are data examples—often derived by making small perturbations to training set examples—that are designed to fool a machine learning model into making an incorrect decision. As an example, (Brown et al., 2017) showed how an object recognition system that could classify an image as containing a banana with high confidence could easily be fooled into making an incorrect classification simply by adding a small sticker of a toaster to the image. One of the key reasons for this brittleness is probably the fact that the model has not “really learned” what a banana is, beyond a collection of pixels arranged in a certain way. We can play such a scenario out to imagine several concerning possibilities in education, for example: a ranking system that places a student in a remedial class because of their test score similarity to a historical batch of remedial students, without factoring in other variables that might better indicate their likelihood of succeeding in more advanced courses (Bergman et al., 2021); facial recognition misclassifications in criminal justice applications that lead to the wrongful incarceration of students or their family members (Hill, 2020); and many more. These scenarios have inspired new directions for building more robust deep learning models (e.g., Tjeng et al., 2019), but the need for awareness about what such models are “doing” in technical terms will remain crucial.

4.2.2. Bias and fairness

Lacking a general understanding of the “how” and “why” behind most decisions, many machine learning models often recapitulate biases in their training data—and hence, risk perpetuating these biases at scale. For example, a recent study illustrated the drastically poor performance of several commercial facial recognition technologies when seeking to identify the faces of black women—due in part to underrepresentation in their training data (Buolamwini & Gebru, 2018). In healthcare, a system for neurological disorder screening based on human speech proved more accurate for individuals who spoke a particular dialect of Canadian English (Gershgorn, 2018). Such shortcomings prevail in education too—with AIED applications favouring certain groups in the content taught, the ways material is covered, and the accuracy of predictions and appropriateness of interventions (Mayfield et al., 2019). The UK’s intention of using predictive models to assign final grades in the wake of 2020 school closures due to the COVID-19 pandemic illustrates this risk: under the modelling scheme, which was eventually dropped, highly-qualified and capable students from historically lower-performing (and lower-resourced) schools were more likely to receive marks lower than what their teachers would have assigned, whereas students in traditionally high-performing private received higher predicted scores (BBC, 2020).

Ultimately, fairness is a highly complex concept, particularly when applied to education (Mayfield et al., 2019); and when and how educationalists choose to use AIED is itself a complex ethical question, even if and when those AIs are optimized to root out bias. Addressing the technical limitations of machine learning will help mitigate the risks outlined above, but it will be insufficient to preempt the full range of educational and ethical issues related to AIED, specifically the application of AI in practice. Multiple significant and important critiques of AIED, and of the use of data in education more broadly, center on issues such as privacy, instrumentalism, surveillance, performance, and governance (Jarke & Berier, 2019; Holmes, et al., 2021; Williamson 2017).

We hope that the technical explanations and considerations outlined in this paper can help inform conversations and decision-making around issues of fair and equitable use—even if they are insufficient to resolve them.

5. Discussion and conclusion

As we have explored, AI is not “one thing”; in this paper we have focused on the more technical aspects of AI to highlight the myriad of (sometimes complementary) computational techniques that collectively constitute AI. Understanding the workings, limitations, and risks associated with each—and especially those powered by machine learning—is critical to developing and deploying them wisely, thoughtfully, and with proper human oversight. Educationalists who do not have a background in computer science or engineering have a vital role to play in this endeavor. To aid with this, we offer the following guiding questions that educationalists may ask as they encounter applications of AI in education, to ensure AI is used ethically, responsibly, and ultimately to improve the human condition:

- **What kind of AI is it?** The examples contained in this paper illustrate how different types of AI can (and cannot) help solve different problems in education, and may help educationalists form a judgement about their applicability and risks within their own contexts. Asking this question may encourage a recognition of both human expertise and the realities of the ‘intelligence’ of AI systems.
- **Does the AI enable something that would be difficult or impossible to achieve without it?** Unpacking any benefits of the scale or diversity of data that the AI operates on, or any efficiencies it enables and weighing them against associated risks or limitations may help justify its usefulness. If an AI-powered system does not enable capabilities or benefits that could be achieved without it, it may not be worth deploying. Just because AI *can* be used to power an education technology system, does not mean it *should* be.
- **What are the potential risks or drawbacks of deploying this technology?** Even in cases where AI might enable high-impact new capabilities, there are likely to be critical failure modes that could lead to unintended, perverse outcomes. Understanding the possibility of, and anticipating, these outcomes is of essential importance.
- **How equitably are the anticipated benefits and risks distributed across different groups of students and families?** AI, especially machine learning-based systems, can “learn,” replicate, and scale bias and inequity. It is therefore important to question whether AI systems might underserve or discriminate against students and families from low-income or minority backgrounds; with disabilities; experiencing varying levels of linguistic proficiency; or facing other vulnerabilities. Asking about past performance or evidence of bias, or about steps taken to ensure equity in application, could be helpful.
- **If you could wave a magic wand and change anything about this technology, what would it be?** All technologies (including those powered by AI) have been designed with a set of values, practices, and use-cases in mind—and therefore, can be changed, even if they appear opaque or difficult to understand. Those who are closest to the application of AI in educational settings should refuse to accept the status quo, using their observations and wisdom to share feedback with system developers in order to spark changes that help improve the human experience with education.

If and how AI should be designed and used in education remains an active question, which can only be answered through conversations between and across different academic communities. As prior work argues, this will require AI researchers and engineers to work with educationalists to better-understand the theory and practice of education. However, we hope we have successfully argued that equally important is the need for educationalists to understand the more technical aspects of theory and practice of AI, especially when critiquing, rejecting or adapting it for their own efforts. Through the provision of an overview of current AI techniques, their use in education, and key limitations and risks, we hope this article will contribute to these on-going conversations and help advance the quest for AIED to improve the human condition.

Acknowledgement

We would like to thank John Hood and Eric Chu for many helpful ideas and comments that helped shape this paper. No human subjects were involved in this research. The authors report no conflicts of interest.

References

- Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An Effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2), 147–179.
- AlphaFold Team. (2020). *AlphaFold: A Solution to a 50-year-old grand challenge in biology*. DeepMind Blog, <https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>
- AlphaStar Team (2019). *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*. DeepMind Blog, <https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., & de Freitas, N. (2016). “Learning to learn by gradient descent by gradient descent.” In *Proceedings of 30th Conference on Neural Information Processing Systems (NIPS)*. <https://proceedings.neurips.cc/paper/2016/file/fb87582825f9d28a8d42c5e5e8b23d-Paper.pdf>
- Ball, S. J. (2017). *The Education debate*. Policy Press.
- Bergman, P., Kopko, E., & Rodriguez, J. E. (2021). Using predictive analytics to track students: Evidence from a seven college experiment. *NBER Working Paper No. 28948*. <https://doi.org/10.3386/w28948>
- Bayyari, M. J., & Berger, J. O. (2004). The Interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1), 58–80.
- BBC (2020). A-Levels and GCSEs: How did the exam algorithm work? *British Broadcasting Company*. <https://www.bbc.com/news/explainers-53807730>
- Bergman, P., & Chen, E.W. (2019). Leveraging parents: The Impact of high-frequency information on student achievement. *Journal of Human Resources*, 56(1), 125–158.
- Bertsimas, D., Delarue, A., & Martin, S. (2019). Optimizing schools’ start time and bus routes. *PNAS*, 116(13), 5943–5948.
- Biesta, G. J. (2015). *Good education in an age of measurement: Ethics, politics, democracy*. Routledge.
- Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. *PsyArXiv*. <https://doi.org/10.48550/arXiv.1712.09665>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research* (Vol. 81, pp. 77–91). <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Castleman, B., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates. *Journal of Economic Behavior & Organization*, 115, 144–160.
- Coleman, C., Baker, R., & Stephenson, S. (2019). A Better cold-start for early prediction of student at-risk status in new school districts. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 732–737). <https://files.eric.ed.gov/fulltext/ED599170.pdf>
- Cook, W. J. (2012). *In pursuit of the traveling salesman: Mathematics at the limits of computation*. Princeton University Press.
- Davis, J. M. V. (2017). The Short and long run impacts of centralized clearinghouses: Evidence from matching teach for America teachers to schools. *Job Market Paper*. <https://ideas.repec.org/p/jmp/jm2017/pda791.html>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. <https://arxiv.org/pdf/1810.04805.pdf>
- Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 153–162). Association for Computational Linguistics.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Neural network modeling and connectionism, No. 10. Rethinking innateness: A Connectionist perspective on development*. The MIT Press.
- Eynon, R. (2013). The Rise of Big Data: What does it mean for education, technology, and media research? *Learning, Media and Technology*, 38(3), 237–240.

- Eynon, R., & Erin Y. (2020). Methodology, legend, and rhetoric: The Constructions of AI by academia, industry, and policy groups for lifelong learning. *Science, Technology, & Human Values*, 46(1), 166-191. <https://doi.org/10.1177/0162243920906475>
- Faria, A.-M., Sorensen, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). *Getting students on track for graduation: Impacts of the Early Warning Intervention and Monitoring System after one year (REL 2017-272)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. <https://files.eric.ed.gov/fulltext/ED573814.pdf>
- Fiacco, J., Cotos, E., & Rosé, C. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 310-319). <https://doi.org/10.1145/3303772.3303808>
- Fischer, C., Parados, Z., & Baker, R. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1), 130-160.
- García-Martín, E., Faviola Rodrigues, C., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134, 75-88.
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64-70.
- Gershgorn, D. (2018). *If AI is going to be the world's doctor, it needs better textbooks*. Quartz. <https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/>
- Goldstein, D. (2019). *San Francisco had an ambitious plan to tackle school segregation. It made it worse*. The New York Times. <https://www.nytimes.com/2019/04/25/us/san-francisco-school-segregation.html>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hakimi, L., Eynon, R., & Murphy, V. A. (2021). The Ethics of using digital trace data in education: A Thematic review of the research landscape. *Review of Educational Research*, 91(5), 671-717. <https://doi.org/10.3102/00346543211020116>
- Hao, K. (2019). The Computing power needed to train AI is now rising seven times faster than ever before. *MIT Technology Review*. <https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/>
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 28(1), 100-108.
- Hennessy, J. L., & Patterson, D. A. (2019). A New golden age for computer architecture. *Communications of the ACM*, 62(2), 48-60.
- Hill, K. (2020). Wrongfully accused by an algorithm. *The New York Times*. <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 504-526. <https://doi.org/10.1007/s40593-021-00239-1>
- Jarke, J., & Breiter, A. (2019). The Datafication of education. *Learning, Media and Technology*, 44(1), 1-6. <https://doi.org/10.1080/17439884.2019.1573833>
- Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 3020-3029). <https://proceedings.mlr.press/v48/johansson16.html>
- J-PAL Evidence Review. (2019). "Will technology transform education for the better?" Abdul Latif Jameel Poverty Action Lab.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A Survey. *Journal of Artificial Intelligence Research*, 4, 237-285.
- Karpicke, J. D., & Roediger III, H. L. (2008). The Critical importance of retrieval for learning. *Science*, 319(5865), 966-968.
- Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Goldstein, D. S. (2013). Estimating the effect of web-based homework. *International Conference on Artificial Intelligence in Education* (pp. 824-827). https://doi.org/10.1007/978-3-642-39112-5_122

- Kim, B., Wattenberg M., Gilmer, J., Cai C., Wexler J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, PMLR* (Vol. 80, pp. 2668-2677). <https://proceedings.mlr.press/v80/kim18d.html>
- Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13), 3521–3526.
- Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *WIREs Cognitive Science*, 6(4), 333–353.
- Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, Vol. 40, E253. <https://doi.org/10.1017/S0140525X16001837>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- Li, F. F., & Etchemendy, J. (2018). *Introducing Stanford's human-centered AI initiative*. <https://hai.stanford.edu/news/introducing-stanfords-human-centered-ai-initiative>
- Luckin, R. (2018). *Machine learning and human intelligence: The Future of education for the 21st century*. UCL IOE Press.
- Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W. (2019). Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 444-460). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4446>
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518, 529–533.
- Murphy, K. (2012). *Machine learning: A Probabilistic perspective*. MIT Press.
- OpenAI Team. (2019). *GPT-2: 1.5B Release*. OpenAI Blog. <https://openai.com/blog/gpt-2-1-5b-release/>
- Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, 3(4). <https://doi.org/10.1177/2332858417749220>
- Pan, S. J., & Yang, Q. (2009). A Survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
- Pathak, P. A., & Shi, P. (2015). How well do structural demand models work? Counterfactual predictions in school choice. *Journal of Econometrics*, 222(1), 161-195. <https://doi.org/10.1016/j.jeconom.2020.07.031>
- Pathak, P. A., & Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *American Economic Review*, 98(4), 1636–1652.
- Perrotta, C., & Selwyn, N. (2020). Deep learning goes to school: toward a relational understanding of AI in education. *Learning, Media and Technology*, 45(3), 251-269.
- Perrotta, C., & Williamson, B. (2018). The Social life of learning analytics: Cluster analysis and the “performance” of algorithmic education. *Learning, Media and Technology*, 43(1), 3-16.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28 (NIPS)* (pp. 505-513). Neural Information Processing Systems.
- Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7). <https://doi.org/10.1126/sciadv.aap7885>
- Prinsloo, P., & Slade, S. (2017). An Elephant in the learning analytics room: the obligation to act. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 46-55). <https://doi.org/10.1145/3027385.3027406>
- Reddy, S., Levine, S., & Dragan, A. (2017). Accelerating human learning with deep reinforcement learning. In *Proceedings of the NIPS'17 Workshop on Teaching Machines, Robots, and Humans* (pp. 1-9). <https://siddharth.io/files/deep-tutor.pdf>
- Ritter, S., Kulikowich, J., Lei, P. W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A Randomized field trial of cognitive tutor Algebra I. *Frontiers in Artificial Intelligence and Applications*, 162(13), 13-20.
- Romero, C., & Ventura, S. (2010). Educational data mining: A Review of the state of the art. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 40(6), 601–618.
- Ross, J., & Macleod, H. (2018). Surveillance, (dis) trust and teaching with plagiarism detection technology. In *Proceedings of the 11th International Conference on Networked Learning* (pp. 235-242). https://www.networkedlearningconference.org.uk/abstracts/papers/ross_25.pdf
- Scharfenberg, D. (2018). *Computers can solve your problem. You may not like the answer*. The Boston Globe. <https://apps.bostonglobe.com/ideas/graphics/2018/09/equity-machine/>

- Shum, S. J. B., & Luckin, R. (2019). Learning analytics and AI: Politics, pedagogy and practices. *British Journal of Educational Technology*, 50(6), 2785-2793.
- Silver, A., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A General reinforcement learning algorithm that masters chess, Shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.
- Sundarajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning, PMLR* (Vol. 70, pp. 3319-3328). <https://proceedings.mlr.press/v70/sundararajan17a.html>
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.
- Tjeng, V., Xiao, K., & Tedrake, R. (2019). Evaluating robustness of neural networks with mixed integer programming. In *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1711.07356>
- Tsai, Y.-S., Perotta, C., & Gašević, D. (2019). Empowering learners with personalised learning approaches? Agency, equity and transparency in the context of learning analytics. *Assessment & Evaluation in Higher Education*, 45(4), 554-567.
- Tuomi, I. (2018). The Impact of artificial intelligence on learning, teaching, and education. *JRC Science for Policy Report, European Commission*. Publications Office of the European Union. <https://doi.org/10.2760/12297>
- Van Hentenryck, P., & Michel, L. (2009). *Constraint-based local search*. MIT Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Williamson, B. (2017). *Big data in education: The Digital future of learning, policy and practice*. SAGE.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057). PMLR.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. <https://doi.org/10.48550/arXiv.1611.01578>

Trends, Research Issues and Applications of Artificial Intelligence in Language Education

Xinyi Huang¹, Di Zou^{2*}, Gary Cheng¹, Xieling Chen¹ and Haoran Xie³

¹Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong

// ²Department of English Language Education, The Education University of Hong Kong, Hong Kong //

³Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR // hxinyi@eduhk.hk // dizoudaisy@gmail.com // chengks@eduhk.hk // xielingchen0708@gmail.com // hrxie2@gmail.com

*Corresponding author

ABSTRACT: Artificial Intelligence (AI) plays an increasingly important role in language education; however, the trends, research issues, and applications of AI in language learning remain largely under-investigated. Accordingly, the present paper, using bibliometric analysis, investigates these issues via a review of 516 papers published between 2000 and 2019, focusing on how AI was integrated into language education. Findings revealed that the frequency of studies on AI-enhanced language education increased over the period. The USA and Arizona State University were the most active country and institution, respectively. The 10 most popular topics were: (1) automated writing evaluation; (2) intelligent tutoring systems (ITS) for reading and writing; (3) automated error detection; (4) computer-mediated communication; (5) personalized systems for language learning; (6) natural language and vocabulary learning; (7) web resources and web-based systems for language learning; (8) ITS for writing in English for specific purposes; (9) intelligent tutoring and assessment systems for pronunciation and speech training; and (10) affective states and emotions. The results also indicated that AI was frequently used to assist students in learning writing, reading, vocabulary, grammar, speaking, and listening. Natural language processing, automated speech recognition, and learner profiling were commonly applied to develop automated writing evaluation, personalized learning, and intelligent tutoring systems.

Keywords: Artificial Intelligence, Language Education, Bibliometric Analysis, Automated Writing Evaluation, Intelligent Tutoring System

1. Introduction

Humans have been steadily improving their learning by employing new technologies. One of the newest technologies in the modern era is Artificial Intelligence (AI), which is defined as “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments” (Organisation for Economic Co-operation and Development, 2019, p. 7). AI has great potential for education as it can generate predictive and diagnostic models for precision education, help visualize students at risk, provide timely intervention, and reduce dropout rates (Lu et al., 2018). Personalized learning systems, software agents, ontologies, and the semantic web are the major AI techniques for education (Hinojo-Lucena et al., 2019). Hwang et al. (2020) categorized AI-powered Education (AIEd) applications into four types. The first type is the intelligent tutor, which can satisfy students’ needs and promote positive learning outcomes. The intelligent tutee is another AIEd application that encourages learners to be tutors and participate in active learning. Intelligent learning tools or partners, the third type, collect and analyze students’ data to enhance learning. The fourth type, policy-making advisor applications, assist administrators in understanding educational trends and problems and help them make effective decisions (Hwang et al., 2020).

Researchers and practitioners of technology-enhanced language learning (TELL) have been applying a wide range of educational technology in language education for three decades (Zou et al., 2018). One of the challenges of using technologies for language learning is that students with different proficiency levels might not achieve the same learning outcomes (Shadiev & Yang, 2020). To solve this problem, machine learning algorithms and data analysis techniques can be used to develop personalized learning systems (Cui et al., 2018). Personalized learning systems allow learners with low language proficiency to learn at their own pace to maximize their progress (Chen et al., 2021a). Heil (2016) observed that many current applications for language learning are decontextualized, lacking authentic speech production. However, AI-enhanced approaches can address this limitation as well. For example, Chen et al. (2019) developed a context-aware ubiquitous language learning system. With a GPS function, this system can support location-based contextualized English learning. The results indicated that students showed high motivation while learning with this AI-enhanced contextualized system and achieved a satisfactory performance. Thus, it appears that AI has great potential for language education and can solve some existing problems and issues in TELL.

The trend of integrating AI into education has hastened the need to analyze AI research. Previous reviews have mainly focused on AI in education in general (e.g., Chen et al., 2020a; Song & Wang, 2020; Zawacki-Richter et al., 2019), while few studies have been conducted examining AI in specific domains, such as language education. To fill this gap, the present study aims to provide a comprehensive review of research on AI-enhanced language learning by noting publication trends, the main research issues, and the most frequently used AI applications in language education during the period 2000-2019. The following research questions guided our study:

- What were the publication trends regarding AI in language education in terms of years, journals, countries, and institutions?
- What were the main research issues in AI-enhanced language education?
- What were the common applications of AI in language education?

2. Literature review

2.1. Review on AI in education

Among all the review studies we found on AI in education, five articles appear to be both the most representative and recent. Chen et al. (2020a), who analyzed the AIED literature from 1999 to 2019, identified a rising frequency of articles in the area, with slow growth between 1999 and 2002, steady growth between 2003 and 2011, and rapid growth between 2012 and 2019. Concerning the key terms used in AIED, “education,” “machine learning,” “robotics,” “artificial intelligence,” and “deep learning” were most frequently used. As for the terms that received growing attention from researchers in recent years, the top ones were “classification,” “STEM” (i.e., science, technology, engineering, and mathematics), “computational thinking,” “educational data mining,” and “neural networks.” Similar results were reported in another study (Chen et al., 2021b) on the past, present, and future of smart learning, an important sub-field of AIED. This review conducted a topic-modeling analysis of 555 relevant articles from 1989 to 2019, identifying several important research issues, including interactive and multimedia learning, STEM education, smart learning analytics, software engineering for e-learning systems, the Internet of Things, and cloud computing.

Zawacki-Richter et al. (2019), who analyzed AI applications in higher education from 2007 to 2019 globally, found that profiling and prediction, assessment and evaluation, adaptive systems and personalization, and ITS were the major AI-enhanced education areas. Regarding profiling and prediction, most studies adopted machine learning methods to model students’ profiles and make predictions. For assessment and evaluation, automated grading systems were frequently used to grade assignments and provide feedback. Adaptive and personalized systems provided academic advice and personalized learning content. ITSs were mainly used to deliver course content and provide learning materials.

Song and Wang (2020) further broadened the scale of analysis by reviewing the development of Educational Artificial Intelligence (EAI) from 2000 to 2019. They proposed that EAI research could be conceptualized as having four stages. The first stage (2000-2004) concentrated on developing intelligent robots, computer programming, and Virtual Reality (VR). A breakthrough in AI occurred during the second stage (2005-2009), with foci on intelligent tutors and educational computing. During the third stage (2010-2014), deep neural networks led to the development of automatic pattern recognition, speech recognition, and image classification. AI infiltrated education at the final stage (2014-2018) when distance education, adaptive learning, e-learning, and data mining became popular.

2.2. Review on AI in language education

Several researchers have conducted reviews on AI in language education. Gamper and Knapp (2002) investigated 40 Intelligent Computer-Assisted Language Learning (ICALL) systems finding that AI techniques such as User Modelling, Natural Language Processing (NLP), Natural Language Generation, Automated Speech Recognition (ASR), and Machine Translation were the most frequently utilized in language learning systems. Ali (2020) reviewed the approaches to integrating AI in language education through content analysis. Ali’s review specifically focused on ASR, which recognizes human speech, identifies linguistic features, and assists in human-machine communication. Related to ASR, Chatbots can conduct intelligent conversations through a keyword matching technique that assesses students’ speaking abilities. AI-amalgamated flipped classrooms can also effectively enhance students’ learning performance and motivation. Therefore, researchers have generally displayed positive attitudes towards AI-enhanced language learning.

Pokrivcakova (2019) analyzed AI technologies from the language teachers' perspective. In the study, different forms of AI were employed in language education for diverse purposes, including: (1) providing personalized learning content; (2) translating a written/spoken text from one language to another; (3) correcting grammar errors by means of writing assistants; (4) conducting conversations using chatbots; (5) creating smart language learning platforms and apps; (6) enabling personalized language tutoring; and (7) developing intelligent VR for learners to practice speaking. Considering the increasing trend of using AI in education, Pokrivcakova (2019) noted the importance of teacher training in the AI age.

Chen et al. (2021a) focused on the sub-field of precision language education and identified research trends and issues in the domain of personalized language learning after reviewing 108 articles between 2000 and 2019. They found that personalized recommendations, feedback, and assessment were the most frequently investigated topics. Findings revealed that personalized language education was effective as it met different learners' needs and provided them with personalized diagnoses and adaptation.

In sum, although many of the existing reviews on AI in education have focused on general education (e.g., Chen et al., 2020a; Song & Wang, 2020; Zawacki-Richter et al., 2019), a few studies have investigated AI in language education; however, most of these have focused on AI tools and applications in language classrooms, with limited research on AI research trends in language education. Moreover, the sample sizes of most previous review studies have been relatively small (e.g., Ali, 2020; Gamper & Knapp, 2002). The wide application of AI in language classrooms suggests AI is playing a significant role in language education, which indicates there is a need to analyze the current research status of AI-enhanced language learning using a computational method that can provide a more comprehensive analysis of the literature. Accordingly, the present study provides an overview of the status of AI in language education by analyzing research trends and the most-discussed topics using bibliometric analysis.

3. Research method

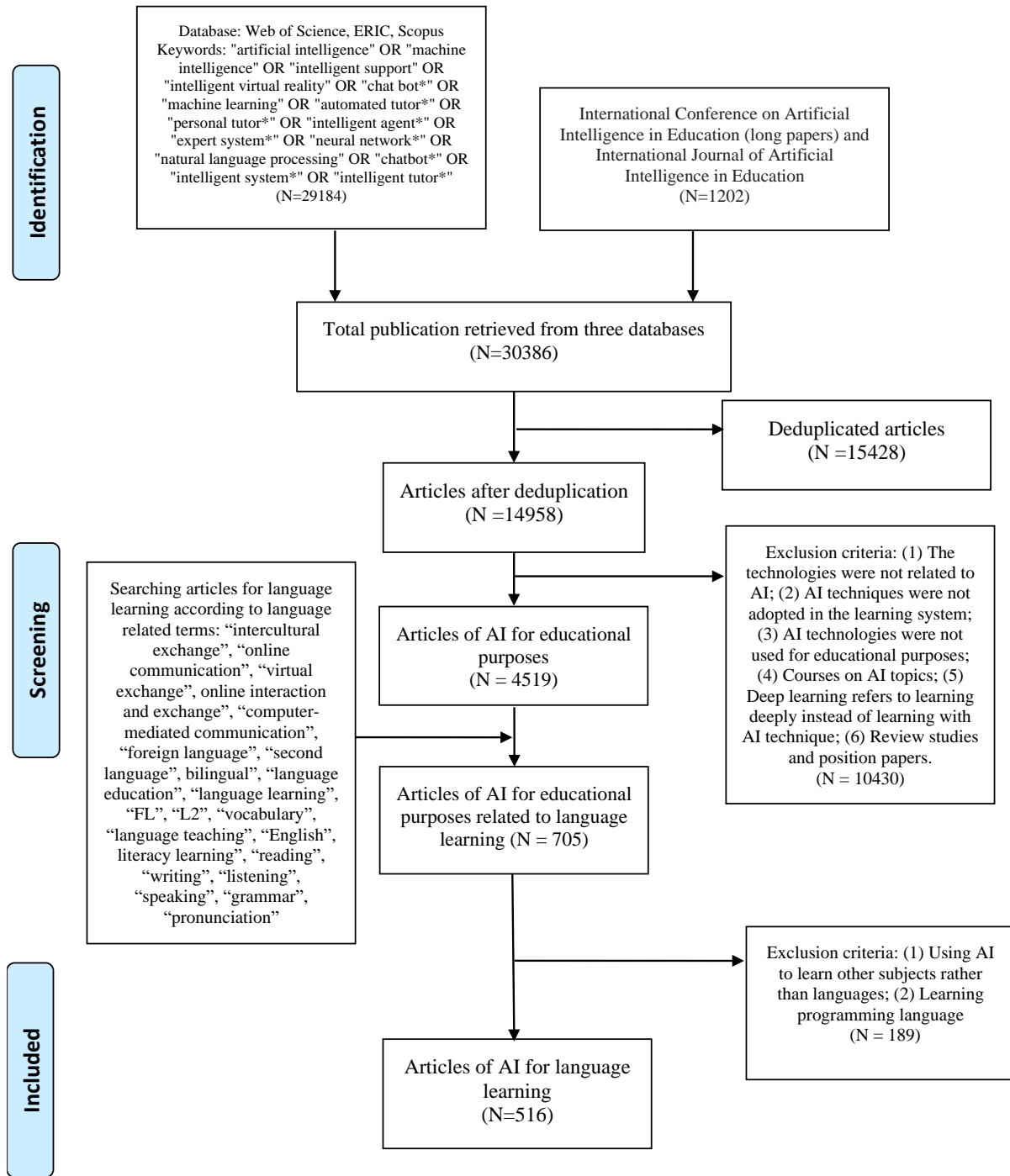
Bibliometric analysis was adopted in this research because it can effectively evaluate the academic status of a particular research area (Chen et al., 2020b; Chen et al., 2020c). Many researchers have employed this method to investigate research trends in different areas. Studies have also applied it to analyze AI in education (e.g., Hinojo-Lucena et al., 2019; Song & Wang, 2020) and language learning (e.g., Chen et al., 2018; Gong et al., 2018). Hence, this method was considered applicable for analyzing AI research trends in language education.

3.1. Data retrieval

Web of Science (WoS), Education Resource Information Center (ERIC), and Scopus were chosen for the databases. Many previous review studies have also included data from these services (e.g., Fu et al., 2022; Wang et al., 2019; Zou et al., 2019; Zou et al., 2020). Following Tran et al. (2019), we selected AI-related keywords in education to search for target papers (Figure 1). A total of 29,184 papers related to AI in education (original research articles) and 1,202 publications from the International Conference on Artificial Intelligence in Education (ICAIE) and the International Journal of Artificial Intelligence in Education (IJAIED) were retrieved. Thus, in total, we identified 30,386 publications. We included conference papers in this review as they are the main source of research on AI in education (Hinojo-Lucena et al., 2019). ICAIE is an important conference in the field, so we also included its proceedings papers to present a comprehensive overview of the whole research area.

Figure 1 illustrates the detailed data retrieval process. After deduplication ($N = 15,428$), two domain experts screened the remaining papers ($N = 14,958$) based on the following criteria: (1) The papers had to focus on AI technologies; (2) AI technologies had to be used to support learning and teaching; and (3) the studies had to be empirical. Upon completion of this initial screening, the inter-coder agreement was 91%, with differences being decided via discussion, resulting in 4,519 remaining papers. We then consulted previous review studies on technology-enhanced language learning and identified 22 language-related keywords (Chen et al., 2021c; Fu et al., 2022; Su & Zou, 2020; Wang et al., 2019; van den Berghe et al., 2019; Zhang & Zou, 2020). Using these keywords (see Figure 1), we searched the titles, abstracts, and keywords of the 4,519 papers and selected those that applied AI for language learning purposes. In total, we found 705 papers. At the final stage, two domain experts examined these papers and excluded those that used AI to learn other subjects or programming languages. The inter-coder agreement was 95%, with differences being resolved via discussion. A total of 516 papers were finalized for review.

Figure 1. Process of data retrieval



3.2. Structural topic modeling

Structural topic modeling (STM) (Roberts et al., 2014) was adopted to identify the latent topics from the 516 papers. STM can identify the principal features of a corpus using machine learning algorithms (Grajzl & Murrell, 2019). We applied this method to extract terms from the titles, abstracts, and keywords. As suggested by Chen et al. (2022), 0.4, 0.4, and 0.2 were respectively assigned as the weights to the terms from keywords, titles, and abstracts. We also employed Term Frequency-Inverse Document Frequencies (TF-IDF) to filter terms according to their importance. Originally, there were 5,582 terms. We set the threshold of TF-IDF as 0.03, 0.04 and 0.05 and found 5,463, 4,935 and 3,807 terms, respectively. We selected terms with 0.04 TF-IDF because 0.03 TF-IDF included terms that were not very relevant (i.e., admissible, diagramming), while 0.05 TF-IDF did not include some important terms (e.g., learn, read). Thus, 0.04 TF-IDF appeared most appropriate. Following previous research (Chen et al., 2020c; Chen et al., 2020d), we ran a set of 16 models by setting the number of topics

ranging from 5 to 20. We compared each model by examining the representative terms and articles according to the following criteria. First, a meaningful topic had to be formed based on the representative terms; second, all articles had to be highly related to the identified topic; third, all topics within a topic model had to be different; and fourth, all crucial dimensions of AI in language education had to be included.

After comparing the 16 models with different numbers of topics, we chose the 10-topic model. We then generated the statistical results based on the level of importance of the topics and obtained the key terms from the topics following the distribution matrix to label the topics. Thereafter, two domain experts interpreted the semantic meanings of each key term and analyzed the representative articles for each topic. Finally, two researchers summarized each topic's labels independently and compared the labeling results to ensure consistency.

3.3. Performance analysis

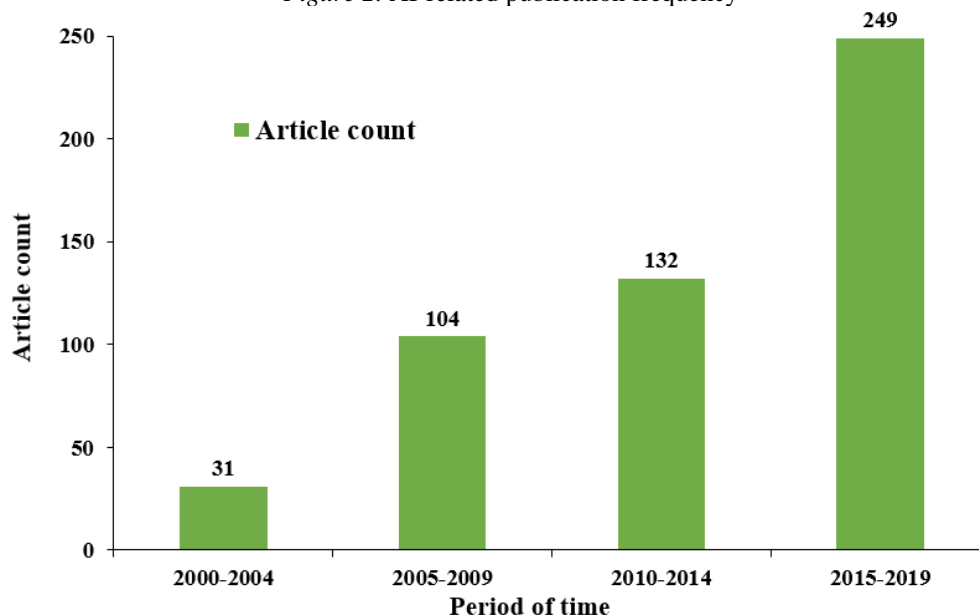
A performance analysis was conducted to investigate the academic outputs of the journals, institutions, and countries/regions. Several indicators were used, including the Hirsch index (H-index), Article count (A), Citation count (C), and Average Citation per Article (ACP). H-index considers both the number of works that have been published as well as the citations of those published papers, which indicates the relevance of the research (Hirsch & Buéla-Casal, 2014) and is one of the most recognized indicators of academic impact (Svensson, 2010). We calculated the citation counts using Google Scholar on 30th May 2020. Google Scholar identifies the most relevant academic information from a given query and offers the citation data, and it is widely considered reliable (Martín-Martín et al., 2018). Many previous review studies have also used the citation counts of Google Scholar for bibliometric analysis (e.g., Chen et al., 2020a; Dey et al., 2018; Wang & Preminger, 2019).

4. Research results

4.1. Publication trends

Figure 2 shows the frequency of articles published on AI-enhanced language learning from 2000 to 2019. A rising trend can be observed, indicating that researchers have paid increasing attention to the field. Researchers paid comparatively little attention to AI-enhanced language learning between 2000 and 2004, while there was a sharp increase during 2005 and 2009. The number of publications kept increasing in the third period (2010-2014) and reached the highest number in the last period (2015-2019).

Figure 2. AI-related publication frequency



4.2. Influential publication sources

Figure 3 presents the top 15 sources that contributed to the research field. The three most influential sources based on the H-index were the *IJAIED*, *Computers & Education*, and *ICAIE*, with H-indexes of 25, 24, and 23 respectively.

ICAIE and *IJAIED* respectively published 264 and 104 articles on AI in language education and accounted for 71% of the total number (516) (see Figure 4).

As for citation counts (Figure 5), the most influential sources were the *ICAIE* (3,294), *IJAIED* (2,540), and *CALICO Journal* (2,234).

As shown in Figure 6, *Bilingualism: Language and Cognition* had the highest ACP (109.92), followed by *Educational Technology & Society* (53.62) and *Language Learning* (52).

Figure 3. Top 15 sources: H-index

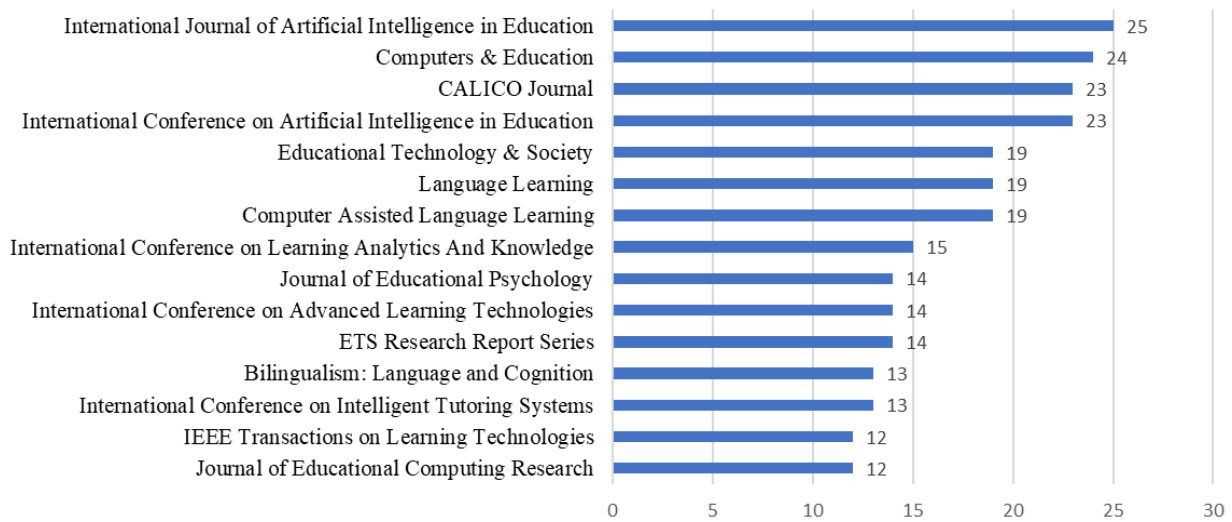


Figure 4. Top 15 publications: Article counts

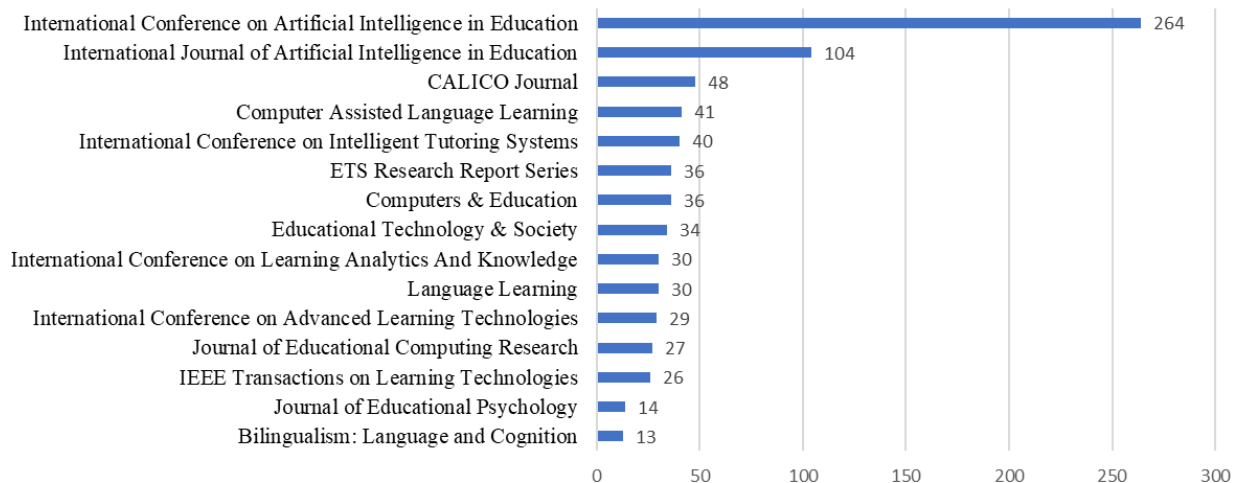


Figure 5. Top 15 publications: Citation counts

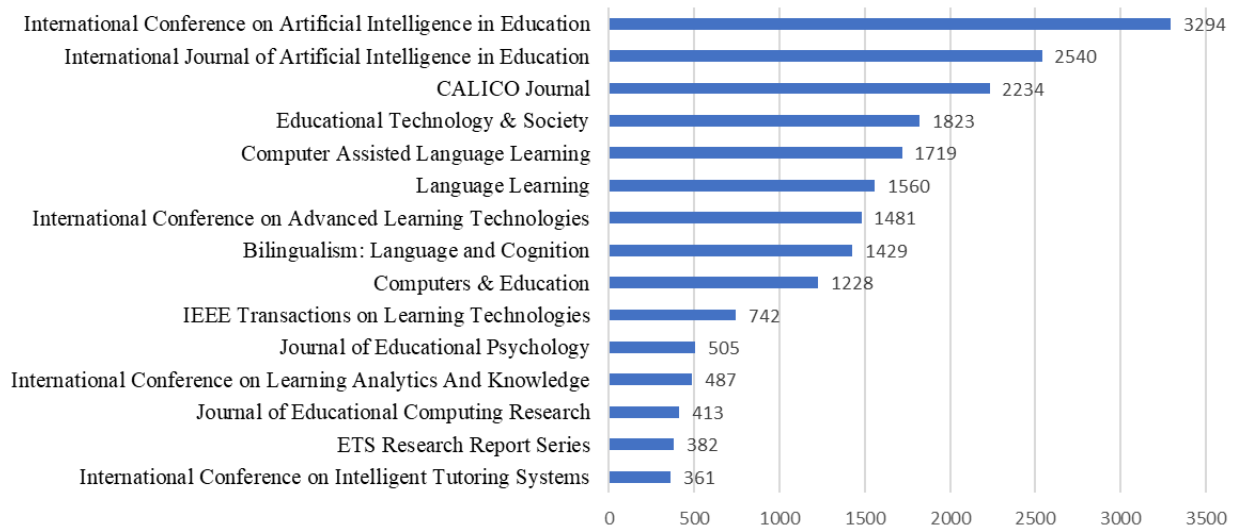
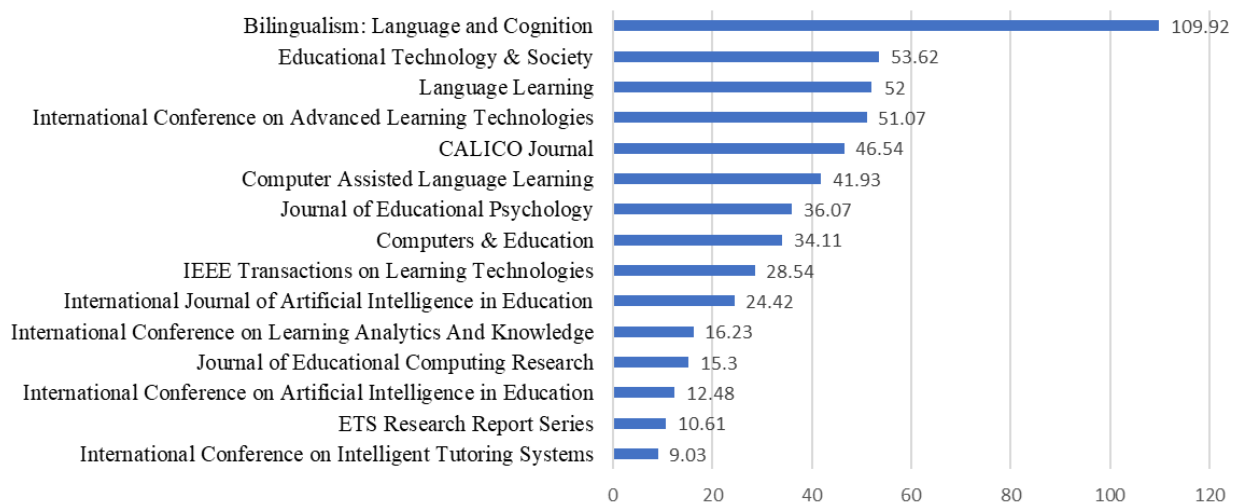


Figure 6. Top 15 publications: Average citation counts



4.3. Representative articles

We also ranked the articles according to their total and annual citations until September 13rd 2020 to identify the most representative studies and analyze their main findings.

- The 10 studies with the highest total citations were Stockwell (2007), Chen et al. (2006), Chen and Li (2010), Johnson et al. (2005), Grimes and Warschauer (2010), Johnson (2007), Calvo et al. (2010), McNamara et al. (2013), Roscoe and McNamara (2013), and McNamara et al. (2015).
- The 10 studies with the highest annual citations were Stockwell (2007), Chen et al. (2006), Chen and Li (2010), McNamara et al. (2015), Alexopoulou et al. (2017), McNamara et al. (2013), Kyle and Crossley (2018), Grimes and Warschauer (2010), Roscoe and McNamara (2013), and Vajjala (2018).

Many studies with the highest total citations also had the highest annual citations, four of which were on AI-enhanced writing. NLP techniques were applied for essay quality evaluation and immediate feedback (Fu et al., 2022). Grimes and Warschauer (2010) investigated teachers' and students' attitudes towards an Automated Writing Evaluation (AWE) tool called MY Access!. This system grades students' essays and provides automated feedback. Results showed that teachers regarded the automatic scoring function as useful because it saved them time, and students considered it helpful for revising and enhancing their writing skills. iWrite was used in Calvo's et al. (2010) study to support collaborative writing activities by helping students revise their group work.

It was found that students spent more time on collaborative writing because the system enabled all group members to view their work, which promoted individual participation. Similarly, McNamara et al. (2013) developed an ITS (Writing Pal) to teach students writing strategies such as generating ideas, organizing essays, and revising essays. This ITS also evaluated essay quality and generated automatic feedback for students. Results indicated that the ratings of this system were similar to that of human graders. Roscoe and McNamara (2013) further examined the feasibility of using this system in writing classrooms. Results from their surveys indicated that students perceived the lessons given by the system as beneficial and informative.

NLP technologies were also used for language feature analysis in AWE and Automated Essay Scoring (AES) systems in the reviewed studies. McNamara et al. (2015) applied a hierarchy classification approach to the AES system that could evaluate essays according to their length and quality and predict scores. Results showed that this approach had a higher accuracy than other AWE systems since it used a set of thresholds to predict essay scores. Alexopoulou et al. (2017) investigated the effects of tasks on learners' written language by analyzing their work using NLP techniques. The results revealed that learners' writing of professional tasks, i.e., writing a job advertisement, had lower error rates than narrative tasks, i.e., storytelling. This is perhaps because professional tasks are normally in bullet-point form. In Kyle and Crossley's (2018) study, NLP was employed to extract language features from the essays of the Test of English as a Foreign Language (TOEFL) and analyze the syntactic complexity of learners' writing. They found that the fine-grained indices of phrasal complexity were the best predictors of learners' writing quality scores because they provided complimentary explanatory power. Similarly, Vajjala (2018) identified the most predictive features in different AES and AWE systems adopting NLP techniques to build predictive models. The researchers concluded that document length played an important role in predicting TOEFL writing scores, and discourse features were an important predictor in Cambridge First Certificate in the English dataset.

In the representative studies, both NLP and ASR were used to enhance communication in game settings. Johnson et al. (2005) integrated AI and serious games into the Tactical Language Training System (TLTS) for language and cultural learning. Learners interacted with the Non-Player Characters (NPC) to complete missions in a simulated world. ASR techniques were used to identify the intended meanings of players' utterances, and NLP was adopted to generate dialogues between the players and the NPCs in the game. However, Johnson et al. (2005) did not evaluate the effectiveness of this game, so it is uncertain whether and to what extent students benefited from learning to use this approach. In a follow-up study, Johnson (2007) evaluated the usefulness of the software by inviting users to rate the system with scores from 0 to 5. Findings revealed that 78% of the participants perceived the training positively and they also felt they had acquired some functional ability of the target language.

Learner profiling, fuzzy item theory, and context-aware techniques have also been integrated into ITS to promote vocabulary learning and reading ability. Stockwell (2007) developed a mobile ITS to enhance students' vocabulary learning. This system keeps logs of students' access to the system, creates learner profiles to record the vocabulary with which students were unfamiliar, and presents these words more frequently. Chen et al. (2016) developed a Personalized Mobile Learning System (PLMS) to recommend English articles to students based on their reading ability. The students' reading ability was evaluated by fuzzy item response theory, and articles were retrieved from websites via a crawler agent. The proposed system was beneficial for students as it provided personalized learning. Chen and Li (2010) designed a personalized context-aware ubiquitous system to provide students with relevant vocabulary learning materials according to their locations, ability, learning time, and leisure time. Results showed that students who applied the learning systems with context awareness outperformed those who did not, as the content was appropriate.

4.4. Productive regions and institutions

Figure 7 lists the top 15 countries/regions ranked by the H-index. The most influential country was the USA (H-index = 38), followed by Taiwan (H-index = 15) and Canada, UK, and Japan (H-index = 11).

Figure 8 shows the USA, Taiwan, and Japan also had the highest citation counts, which were 5,808, 1,333, and 678, respectively.

Figure 9 shows the USA ($n = 228$), Japan ($n = 44$), and Taiwan ($n = 39$) produced the greatest number of studies with the USA contributing 44% of the total publications.

New Zealand had the highest ACP (37.14), followed by Taiwan (34.18) and Hong Kong (32.88) (Figure 10).

Figure 7. Top 15 countries/regions: H-index
H-index

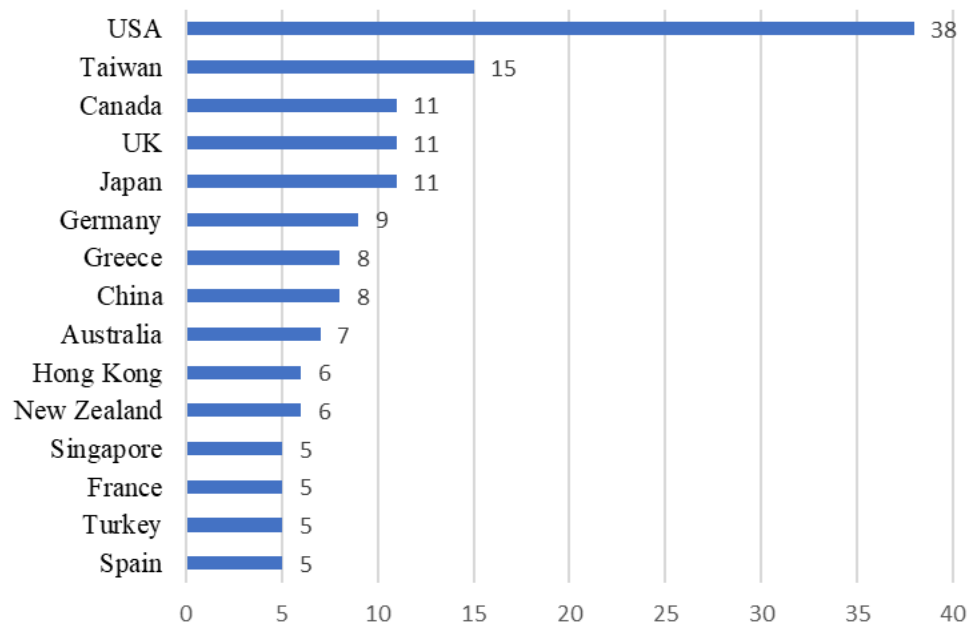


Figure 8. Top 15 countries/regions: Citation counts
Citation counts

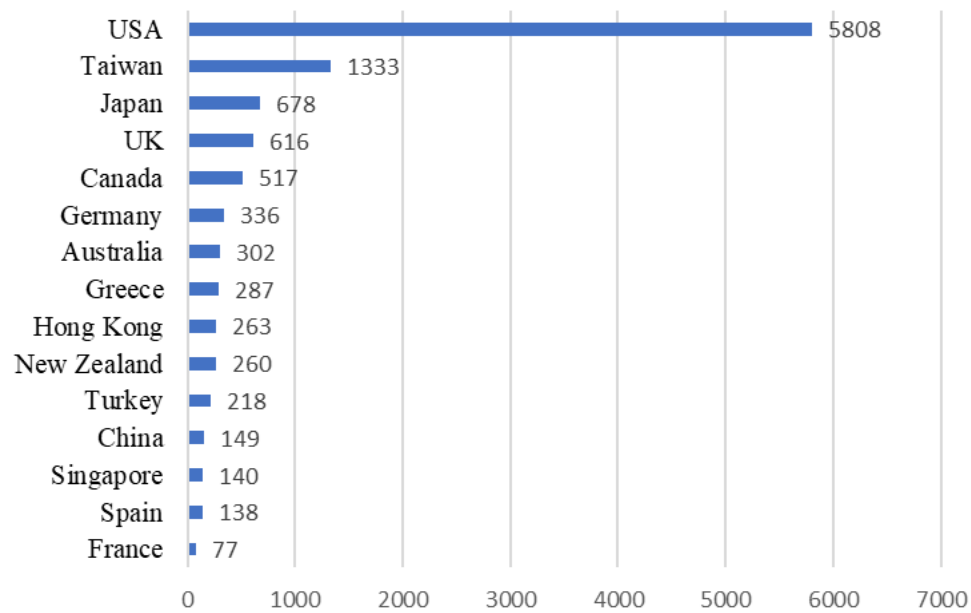


Figure 9. Top 15 countries/regions: Article counts

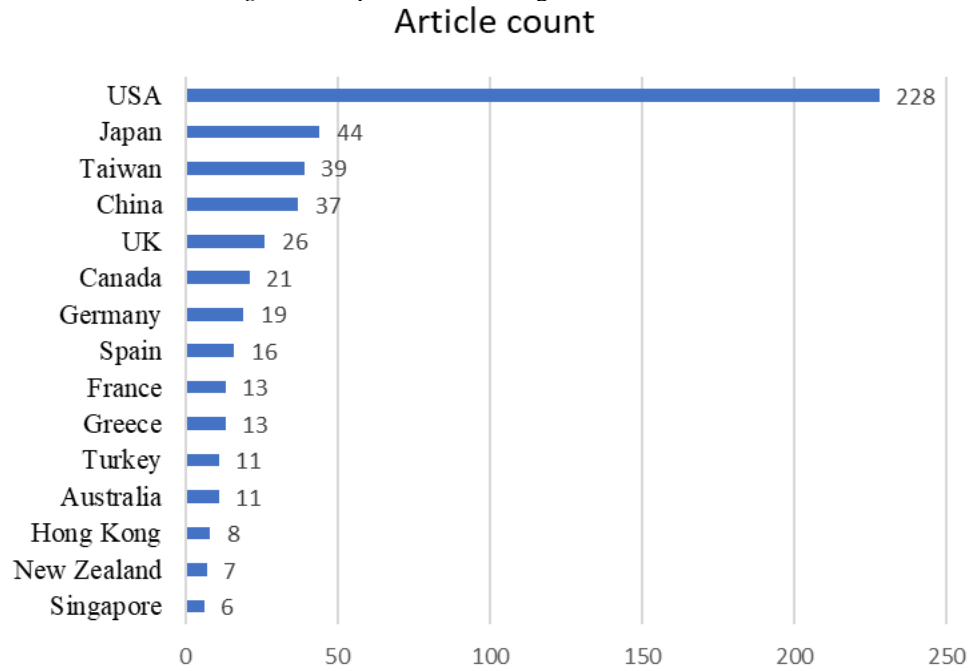
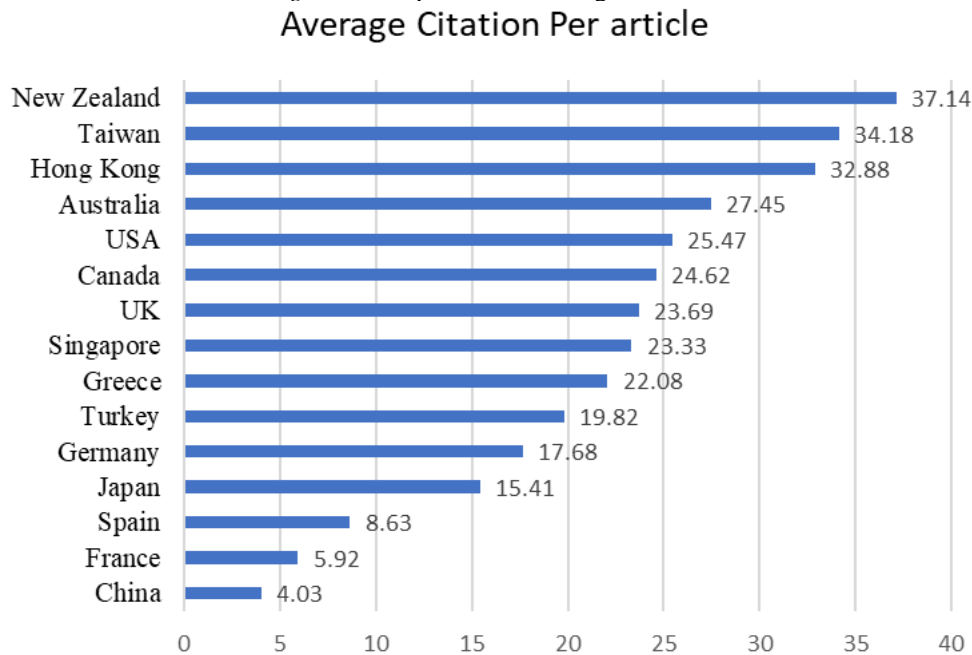


Figure 10. Top 15 countries/regions: ACP



The USA had the highest H-index and the greatest article and citation counts, and 10 out of the 13 top institutions ranked by H-index were from the USA (Figure 11). The other institutions were in Greece (*University of Piraeus*) which ranked seventh, Germany (*University of Tübingen*) which ranked ninth, and Australia (*University of Technology Sydney*) which ranked 13th.

The top three institutions were *Arizona State University* (H-index = 17), *Georgia State University* (H-index = 13), and *Carnegie Mellon University* (H-index = 13). These three also had the greatest article counts, which were 56, 29, and 27, respectively (Figure 12).

Arizona State University was the university that had their papers most frequently cited by researchers ($n = 1,093$), followed by *Pennsylvania State University* ($n = 910$) and *Georgia State University* ($n = 800$) (Figure 13).

The universities with the highest ACP were *Pennsylvania State University* ($n = 82.73$), *University of Southern California* ($n = 63.44$), and *University of Pittsburgh* ($n = 41.54$) (Figure 14).

Figure 11. Top 13 institutions: H-index
H-index

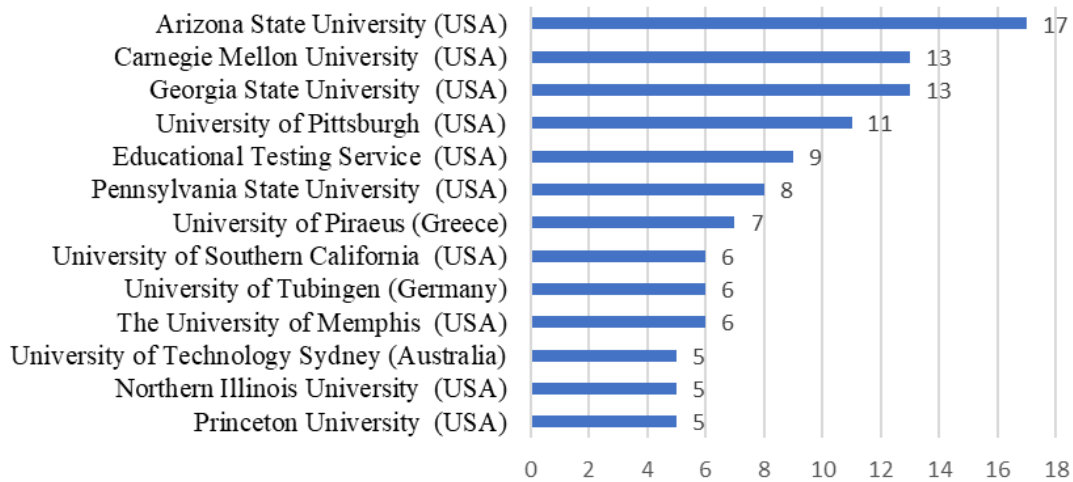


Figure 12. Top 13 institutions: Article count
Article counts

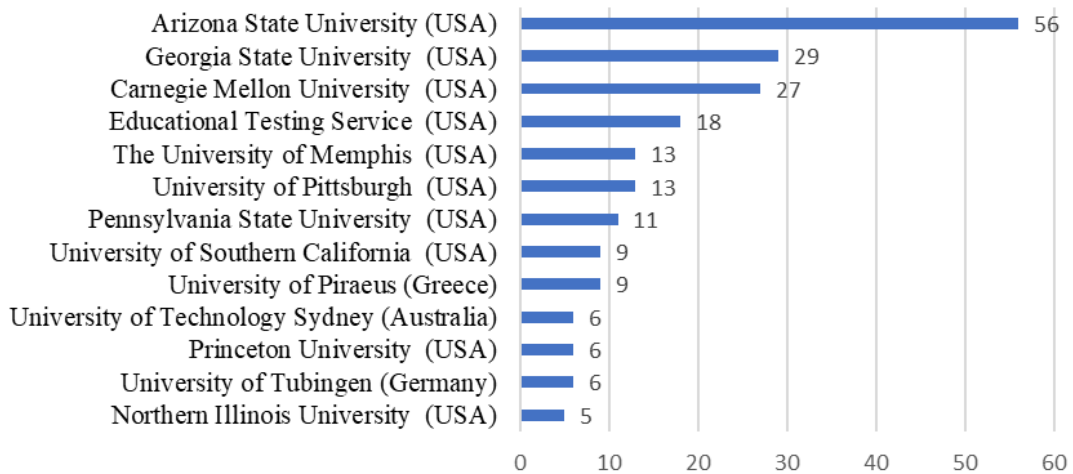


Figure 13. Top 13 institutions: Citation count
Citation counts

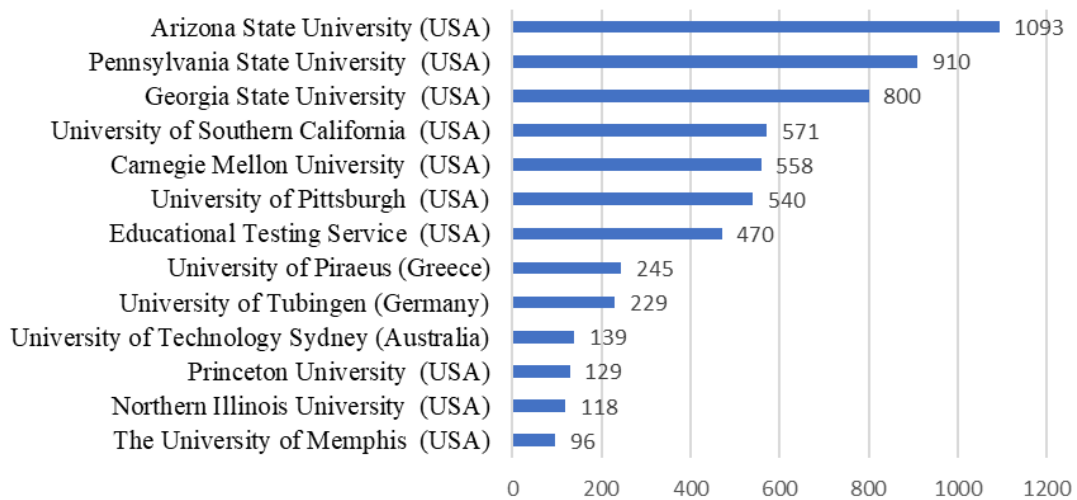
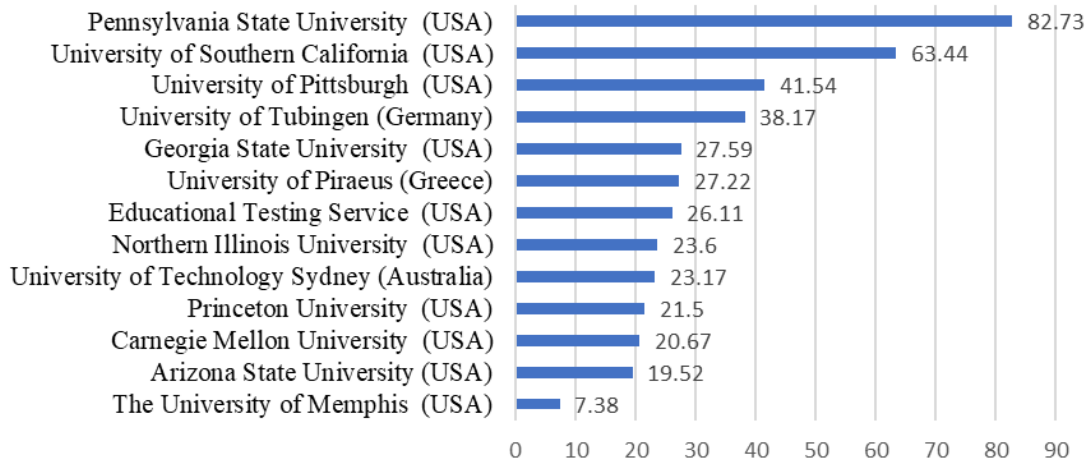


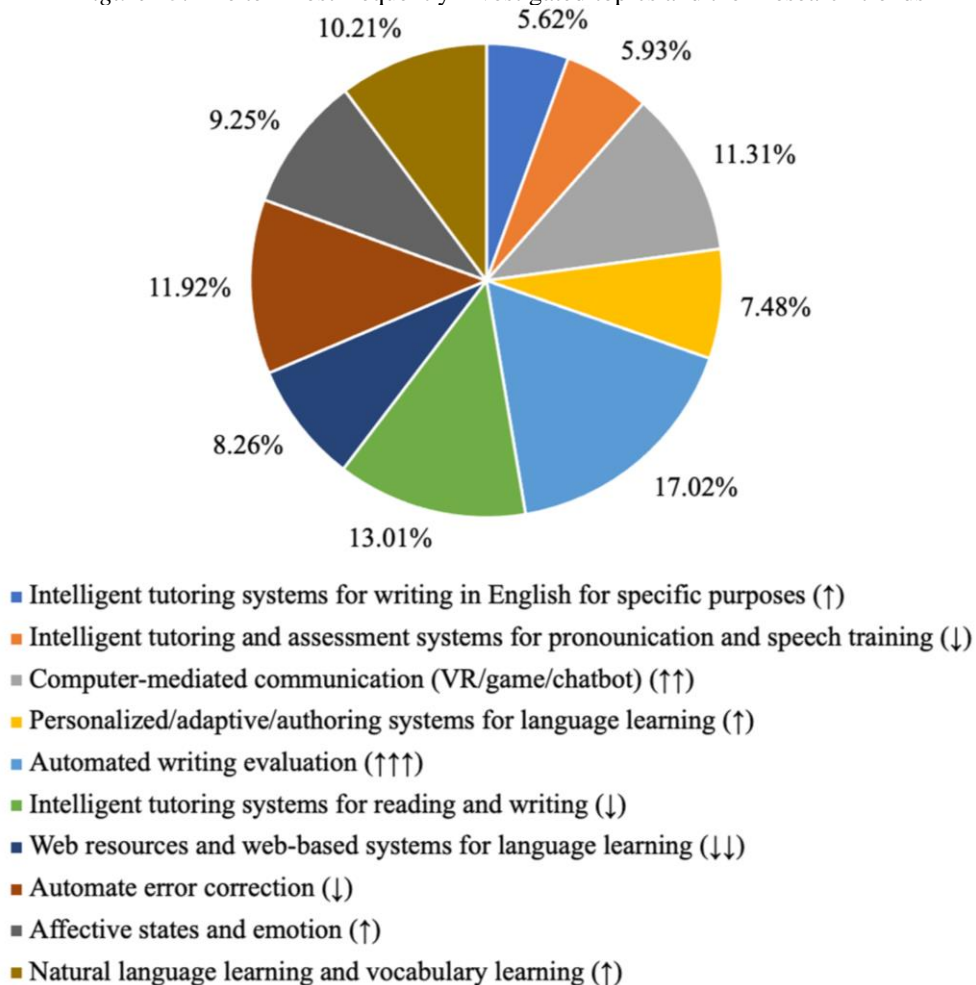
Figure 14. Top 13 institutions: ACP
Average Citation Per article



4.5. Research foci and trends

Figure 15 presents the 10 most frequently investigated topics in AI-assisted language learning and their research trends (indicated by arrows).

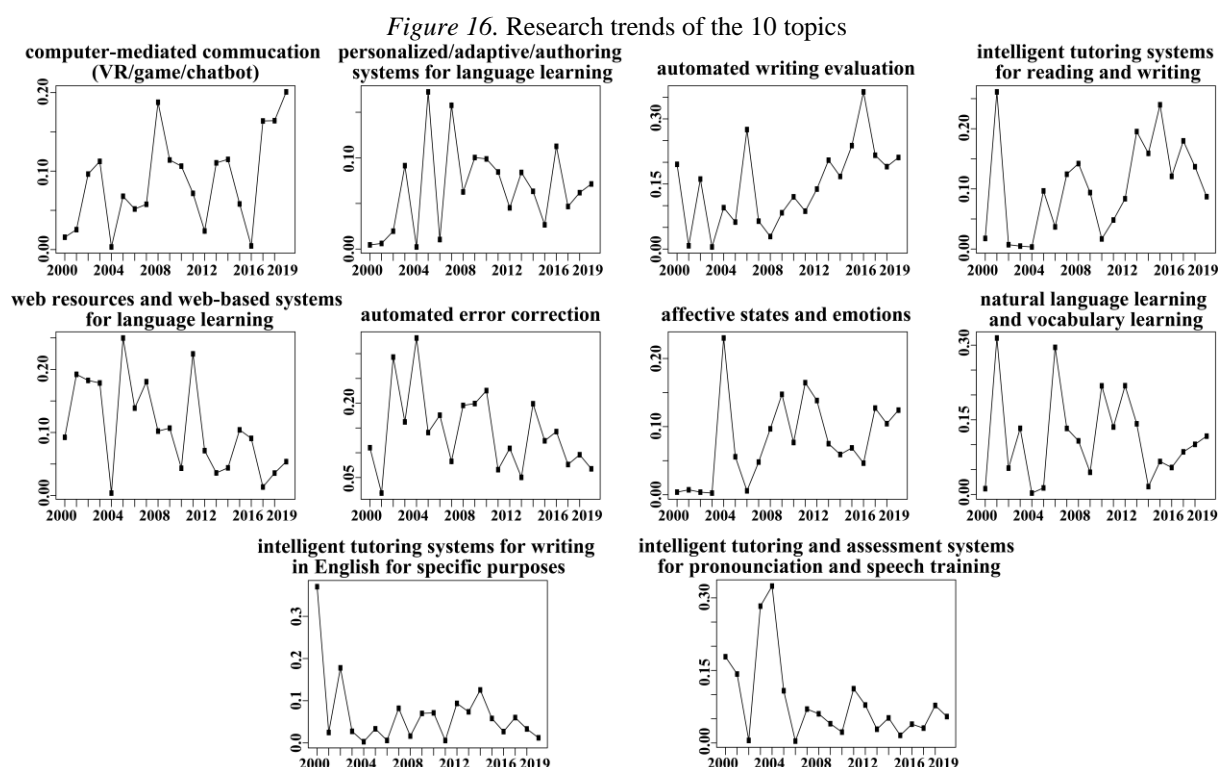
Figure 15. The ten most frequently investigated topics and their research trends



Note. Arrows represent increasing (up) and decreasing (down) interest over time with intensity indicated by the number of arrows.

AWE was the most popular topic accounting for 17.02% of the articles. ITS for reading and writing was comparatively less explored at 13.01%. Automated Error Detection and CMC shared a similar proportion at 11.92% and 11.31%, respectively. The fifth major issue was Natural Language Learning and Vocabulary Learning, which accounted for 10.21% of the reviewed articles. Most of these topics continued to attract research interest towards the end of the review period apart from four topics, i.e., Web resources and Web-based systems for language learning, Automated Error Detection, ITS for writing in English for Specific Purposes, and ITS and Assessment System for pronunciation and speech training. The topics, Web Resources and Web-based systems for language learning, exhibited a significant decreasing trend.

Figure 16 presents the annual article counts of each topic. Both CMC and AWE significantly increased in the later years. Studies on CMC generally increased with fluctuations throughout the period having the smallest number of articles between 2004 and 2016 but having the most in 2019. From 2008, the number of articles on AWE increased and achieved its highest point in 2016 but dropped in the final three years. Web resources and Web-based systems for language learning gradually decreased with a sharp decrease in 2004; in later years, it fluctuated.



5. Discussion

5.1. Research trends of AI in language learning

The results reveal that the number of articles on AI in language learning increased greatly in 2016. Zawacki-Richter et al. (2019), Chen et al. (2020a), and Chen et al. (2022) also showed that research on AI in education increased greatly from 2016, based on which, AI applications developed rapidly. Concerning the number of articles published in each journal, *ICAIE* had the greatest number from 2000 to 2019, followed by *IJAIED*, which is generally consistent with Zawacki-Richter et al. (2019); however, they found that *IJAIED* contributed the most articles from 2007 to 2019. This is likely because we also included conference papers in our review.

Similar to the review findings of Song and Wang (2020) and Zawacki-Richter et al. (2019), we found that the USA had the most publications from 2000 to 2019. However, while Song and Wang (2020) found that the UK and China ranked second and third, respectively, our review ranked Japan and Taiwan in these positions. This is likely because our research focused mainly on AI in language education, whereas they investigated AI in education in general.

Our review indicated that AWE systems were frequently used in language learning, given their potential to reduce teachers' workload and assist students in writing and revision. Additionally, AI may be integrated into VR technology to help students practice their target language in simulated environments (Mirzaei et al., 2018). This innovative approach to CMC drew increasing attention in the later years of our review. With the rising number of ITSs, conventional web-based learning systems drew less research interest in the later years. Similar findings were also reported in Johnson et al. (2017).

5.2. Common applications of AI in language learning

5.2.1. AI applications in learning writing

AI was used to assist students' writing via AWE systems and ITS. These systems evaluate students' work using NLP techniques to diagnose and comment on students' errors so that they have a comprehensive understanding of language use. In Lee et al. (2015), a correction system called Genie Tutor was designed to improve English writing by detecting grammar mistakes and suggesting appropriate expressions. This system guided learners to correct their mistakes in real-time, which is useful for language development. An ITS (i.e., EJP-Write) was also developed to facilitate academic journal writing in Lin et al. (2017). The system's functions, such as citing references and searching for templates, were helpful and effective in providing students with phrase and paragraph templates for better language use.

5.2.2. AI applications in learning reading

ITS was also used to enhance language learners' reading comprehension. For example, Johnson et al. (2017) developed an ITS, Interactive Strategy Training for Active Reading and Thinking (iSTART), for adult literacy learners. iSTART offered instructional videos and exercises for learning comprehension strategies. It also taught summarization strategies and provided learners with interactive narratives to read. The results indicated that learners had positive attitudes towards the narratives. Another example was Wijekumar et al. (2017), who developed an ITS to teach Structure Strategy (ITSS) for enhancing reading comprehension. The ITSS helped students identify text structures and provided hints and feedback in an assessment exercise. The results indicated that the students who used the ITS outperformed those who did not as the system helped organize textual information.

5.2.3. AI applications in learning vocabulary and grammar

One example of AI for vocabulary learning was a study by Chen and Li (2010), who developed a context-aware vocabulary learning system. The system could also suggest new words to be learned based on students' leisure time, i.e., new words would be suggested if the students had more time to learn. Results showed that the students who used the system with context awareness outperformed those who did not. In another study, Pandarova et al. (2019) developed an ITS for practicing English tenses. This system applied dynamic difficulty adaption to adjust the difficulty levels of grammar exercises. The results showed that the system could provide materials of appropriate difficulty levels and allow students to learn grammar at their own pace, making it conducive to effective learning.

5.2.4. AI applications in learning speaking and listening

AI was frequently used to facilitate speaking and listening. Ayedoun et al. (2019) developed a conversational agent to foster communication. The agent was designed based on communication strategies and affective backchannels. The learners could practice and improve their conversation skills by asking the AI agent questions which were then answered. In Johnson (2007), learners practiced speaking skills in games, i.e., the Mission and Arcade. While playing the Arcade Game, players are required to give spoken commands to move their avatars, and in the Mission game, the players speak on behalf of their avatars to complete their mission. ASR techniques were embedded in the games to enable learners to interact with the NPC to practice speaking and listening. The results showed most participants felt the game helped them acquire functional abilities in the target language.

5.3. Advantages of using AI in language learning

5.3.1. Providing personalized learning experiences

AI can suggest appropriate content for learners according to their level, needs, and preferences with advanced algorithms. Pandarova's et al. (2019) system could adjust the difficulty of grammar learning content according to students' language abilities, which allowed students to learn at their own pace, optimizing the learning outcomes. Similarly, Chen et al. (2006) designed a PIMS to enhance reading development. The system recommended English news articles based on a learner's language proficiency. Results showed that using this personalized system for facilitating students' reading was effective as it reduced cognitive overload by aligning articles with students' level of competence. In Chao et al. (2012), the Affective Tutoring System recommended lessons based on learners' emotional state. The system monitored students' moods and customized learning materials to help students avoid learning anxiety. If the system detected negative emotions, it provided relatively easier learning tasks. In this way, students' self-confidence was enhanced, thereby encouraging them to learn.

5.3.2. Enabling immediate adjustment

AI enabled language learners to adjust their learning after receiving automated feedback. As discussed, the NLP techniques used in the AWE systems can detect errors and provide learners with rich feedback, which allows them to take immediate action. For example, the Bengali Handwriting Education System used in Khatun and Miwa (2016) recognized learners' errors such as stroke production errors and stroke sequence errors. Students received timely feedback and made immediate adjustments using this system. In this way, students' language proficiency could be enhanced by repeatedly making modifications and improving their work. As for the quality of feedback, Gierl et al. (2014) showed that AWE systems can provide rich formative feedback, which can overcome teachers' preference for summative feedback due to time constraints with large-sized classes. Gierl et al. (2014) offered students AI-based rich and individualized feedback, enabling them to adjust their learning behavior during their learning process instead of at the final stage.

5.3.3. Rich opportunities using AI in language learning

Using AI techniques, the limited opportunities to practice the target language can be resolved. ITS allows students to learn anywhere and anytime. Stockwell (2007) developed a mobile-based ITS that could record difficult words by presenting them more frequently to increase learning opportunities. Learners could also practice the target language by interacting with a digital human. Mirzaei et al. (2018) introduced Virtual Reality Conversation Envisioning for learners to interact with an AI agent in an immersive context under which simulated scenarios, e.g., bargaining and interviewing, could be created. Students had more opportunities to practice their speaking skills by conducting conversations in different contexts and had more frequent use of the language without going abroad.

5.4. Challenges using AI in language learning

5.4.1. Reliability of AI technology

Although we have discussed the effectiveness of applying AI in language learning in previous sections, its reliability remains a concern. Many researchers have expressed uncertainty about whether this technology is ready for use in the classroom. Grimes and Warschauer (2010) doubted the accuracy of AWE as it could not evaluate subjective features of natural languages. The computational semantic analysis mainly focuses on the denotative meanings of words, while the connotative meanings may not be fully captured. In such cases, the author's intent is unlikely to be evaluated by the system resulting in improper grading of essays. Similarly, Johnson (2007) noted the challenges of evaluating ASR accuracy. Since ASR performance varies across contexts, students may not have smooth interactions with the NPC. As the quality of interactions directly impacts students' learning effectiveness, uncertainty regarding quality can pose challenges for using AI to learn languages. More advanced AI technology is needed to address this problem, and system developers should extensively test their designs before launching new systems.

5.4.2. Acceptance by teachers and students

The uncertain effectiveness of using AI in language education is sometimes caused by teachers' and students' reluctance to use the technology. For example, students in Roscoe and McNamara's study (2013) complained that some feedback given by the writing system was confusing. As the quality of AI cannot be guaranteed, students and teachers may have little motivation to use it as prior negative experiences in using technology can discourage them. Lin et al. (2017) found that users who had little experience using e-learning tools had lower satisfaction with ITS and had negative perceptions of the system due to its differences from traditional technology. Such challenges were also noticed by Pokrivcakova (2019), who showed that a lack of experience with Information Communication Technology (ICT) resulted in teachers' reluctance to use AI-related technologies. Thus, the acceptance of instructors and learners could be improved by developing better AI-enhanced learning systems that provide better teaching and learning experiences and help build positive attitudes. Teacher training programs should also be conducted to help teachers understand the potential benefits of AI in language education.

5.4.3. Social issues of AI in language education

Discourse analysis conducted by AI may be biased if the data and algorithms used for training contain societal biases (Yang et al., 2021). Algorithms may include unbalanced and disproportionate information (Luan et al., 2020) which could lead to social inequities or social cohesion. Further, as some developing countries cannot afford basic ICTs, they may be unlikely to adopt newly developed AI-based technologies possibly leading to a more expansive digital divide and contributing to educational inequality (Luan et al., 2020). Hwang et al. (2020) and Zhang and Aslan (2021) have also suggested that AIED ethics be developed to address privacy issues from all stakeholders. For example, principles and ethical codes could be established before using AI to avoid leaking personal information. Researchers need to screen out biased information or set up keywords to filter sensitive information when selecting resources for natural language processing. International organizations could also support developing countries by providing essential communication technologies (Luan et al., 2020).

Putting humans at the center of AI applications is an important consideration. AI needs to be shifted from technology-oriented applications, which emphasize the development of production and performance, to human-oriented ones, which accentuate the integration of human and machine intelligence (Yang, 2021; Yang et al., 2021). Yang et al. (2021) called this new trend of AI, Human-centered AI (HAI), suggesting that some of the limitations of AIED can be solved by HAI. HAI algorithms such as Bidirectional Encoder Representations from Transformers and Generative Pre-Training can be adopted for natural language processing to achieve performances close to those of humans (Yang et al., 2021). This can help increase the accuracy of grading on student writing. HAI also allows researchers to understand users' perceptions and requirements when using AI-enhanced language tools (e.g., translation applications and voice assistants). It can identify students' motivations and engagement and provide them with timely assistance and intervention during the learning process, which is essential for effective language learning (Huang et al., 2020).

6. Conclusion

The present review provides comprehensive coverage AI research trends in language education by analyzing publications from 2000 to 2019. According to our results, the number of articles related to AI in language education showed an increasing trend over the period reflecting researchers' growing interest in using AI tools to assist language learning. Notably, an increasing number of new journals on AI, such as *Computers & Education*, *Artificial Intelligence*, *International Journal of Learning Analytics* and *Artificial Intelligence for Education*, and *IJAIED* emerged during the period. *IJAIED* was the most influential journal, and the USA and the Arizona State University were the country and institution that contributed the most research. We also found that AWE is the most investigated AI application, and its interest grew over the years.

As for the limitations of our review, because it was limited to articles found in only three sources (i.e., WoS, ERIC, and Scopus), not all academic research related to AI in language education was included. Thus, future research may consider including more sources to provide a more comprehensive analysis. Regarding the citation count, the data retrieved from Google Scholar might have included citations from non-academic resources, which may have led to multiple counts when the publications were released on different platforms. Future research may consider using other approaches for citation counting. Additionally, the research methodology applied in the current research was bibliometric; future research may apply different methods to further

investigate the literature on AI in language education from other perspectives. Other suggestions for future research on AI in language education include Yang et al. (2021), who recommended investigating AI's potential for improving teaching and learning outcomes, and Hwang et al. (2020) who also suggested that future research investigate the possibility of using AI for language courses.

Acknowledgement

An abstract entitled "Artificial Intelligence in Language Education" based on this paper was presented at the International Conference on Education and Artificial Intelligence 2020, The Education University of Hong Kong, 9-11 November 2020, Hong Kong. Gary Cheng's work in this research is supported by the Research Cluster Fund (RG 78/2019-2020R) of The Education University of Hong Kong and the Dean's Research Fund 2019/20 (IDS-2 2020) of The Education University of Hong Kong. Haoran Xie's work in this research is supported by the Faculty Research Fund (DB21A9) and the Lam Woo Research Fund (LWI20011) of Lingnan University, Hong Kong.

References

- Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A Large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180-208.
- Ali, Z. (2020). Artificial Intelligence (AI): A Review of its uses in language teaching and learning. *IOP Conference Series: Materials Science and Engineering*, 769(1), 012043. <https://doi.org/10.1088/1757-899x/769/1/012043>
- Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education*, 29(1), 29-57.
- Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2010). Collaborative writing support tools on the cloud. *IEEE Transactions on Learning Technologies*, 4(1), 88-97.
- Chao, C. J., Lin, H. K., Huang, T. C., Hsu, K. C. & Hsieh, C. Y. (2012). The Application of affective tutoring systems (ATS) in enhancing learners' motivation. In *Workshop Proceedings of the 20th International Conference on Computers in Education (ICCE)* (pp. 58-66). Asia-Pacific Society for Computers in Education.
- Chen, C. M., Hsu, S. H., Li, Y. L., & Peng, C. J. (2006). Personalized intelligent m-learning system for supporting effective English learning. In *2006 IEEE International Conference on Systems, Man and Cybernetics* (Vol. 6, pp. 4898-4903). IEEE. <https://doi.org/10.1109/ICSMC.2006.385081>
- Chen, C. M., & Li, Y. L. (2010). Personalised context-aware ubiquitous learning system for supporting effective English vocabulary learning. *Interactive Learning Environments*, 18(4), 341-364.
- Chen, M. P., Wang, L. C., Zou, D., Lin, S. Y., & Xie, H. (2019). Effects of caption and gender on junior high students' EFL learning from iMap-enhanced contextualized learning. *Computers & Education*, 140, 103602. <https://doi.org/10.1016/j.compedu.2019.103602>
- Chen, X., Hao, J., Chen, J., Hua, S., & Hao, T. (2018). A bibliometric analysis of the research status of the technology enhanced language learning. In *International Symposium on Emerging Technologies for Education* (pp. 169-179). Springer, Cham.
- Chen, X., Xie, H., & Hwang, G. J. (2020a). A Multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, 1, 100005. <https://doi.org/10.1016/j.caeai.2020.100005>
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020b). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- Chen, X., Zou, D., & Xie, H. (2020c). Fifty years of British Journal of Educational Technology: A Topic modeling based bibliometric perspective. *British Journal of Educational Technology*, 51(3), 692-708.
- Chen, X., Zou, D., Cheng, G., & Xie, H. (2020d). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A Retrospective of all volumes of computer & education. *Computers & Education*, 103855. <https://doi.org/10.1016/j.compedu.2020.103855>
- Chen, X., Zou, D., Xie, H., & Cheng, G. (2021a). Twenty years of personalized language learning. *Educational Technology & Society*, 24(1), 205-222.
- Chen, X., Zou, D., Xie, H., & Wang, F. L. (2021b). Past, present, and future of smart learning: A Topic-based bibliometric analysis. *International Journal of Educational Technology in Higher Education*, 18(1), 1-29.

- Chen, X., Zou, D., Xie, H. R., & Su, F. (2021c). Twenty-five years of computer-assisted language learning: A Topic modeling analysis. *Language Learning & Technology*, 25(3), 151-185.
- Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of Artificial Intelligence in education. *Educational Technology & Society*, 25(1), 28-47.
- Dey, A., Billinghamurst, M., Lindeman, R. W., & Swan, J. (2018). A Systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI*, 5, 37. <https://doi.org/10.3389/frobt.2018.00037>
- Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2022). A Review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, 1-43. <https://doi.org/10.1080/09588221.2022.2033787>
- Gamper, J., & Knapp, J. (2002). A Review of intelligent CALL systems. *Computer Assisted Language Learning*, 15(4), 329-342.
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48(10), 950-962.
- Gong, Y., Lyu, B., & Gao, X. (2018). Research on teaching Chinese as a second or foreign language in and outside mainland China: A Bibliometric analysis. *The Asia-Pacific Education Researcher*, 27(4), 277-289.
- Grajzl, P., & Murrell, P. (2019). Toward understanding 17th century English culture: A Structural topic model of Francis Bacon's ideas. *Journal of Comparative Economics*, 47(1), 111-135.
- Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A Multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, 8(6), 1-44. <https://ejournals.bc.edu/index.php/jtla/article/view/1625>
- Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A Review of mobile language learning applications: Trends, challenges, and opportunities. *EuroCALL Review*, 24(2), 32-50. <https://doi.org/10.4995/eurocall.2016.6402>
- Hinojo-Lucena, F. J., Aznar-Díaz, I., Cáceres-Reche, M. P., & Romero-Rodríguez, J. M. (2019). Artificial intelligence in higher education: A Bibliometric study on its impact in the scientific literature. *Education Sciences*, 9(1), 1-9. <https://doi.org/10.3390/educsci9010051>
- Hirsch, J. E. & Bucla-Casal, G. (2014). The Meaning of the H-index. *International Journal of Clinical and Health Psychology*, 14(2), 161-164.
- Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: Evaluation of classification methods and learning logs. *Interactive Learning Environments*, 28(2), 206-230.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- Johnson, A. M., Guerrero, T. A., Tighe, E. L., & McNamara, D. S. (2017). iSTART-ALL: Confronting adult low literacy with intelligent tutoring for reading comprehension. In *International Conference on Artificial Intelligence in Education* (pp. 125-136). Springer. https://doi.org/10.1007/978-3-319-61425-0_1
- Johnson, W. L. (2007). Serious use of a serious game for language learning. *Frontiers in Artificial Intelligence and Applications*, 158, 67-74.
- Johnson, W. L., Vilhjálmsdóttir, H. H., & Marsella, S. (2005). Serious games for language learning: How much game, how much AI? In *Artificial Intelligence in Education* (pp. 306-313). IOS Press.
- Khatun, N., & Miwa, J. (2016). An Autonomous learning system of Bengali characters using web-based intelligent handwriting recognition. *Journal of Education and Learning*, 5(3), 122-138.
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333-349.
- Lee, K., Kwon, O. W., Kim, Y. K., & Lee, Y. (2015). A Hybrid approach for correcting grammatical errors. In F. Helm, L. Bradley, M. Guarda, & S. Thounesny (Eds.), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 362-367). Research-publishing.net. <https://doi.org/10.14705/rpnet.2015.000359>
- Lin, C. C., Liu, G. Z., & Wang, T. I. (2017). Development and usability test of an e-learning tool for engineering graduates to develop academic writing in English: A Case study. *Educational Technology & Society*, 20(4), 148-161.
- Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C. C. (2020). Challenges and future directions of big data and artificial intelligence in education. *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.580820>
- McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2), 499-515.

- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A Hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35-59.
- Mirzaei, M. S., Zhang, Q., van der Struijk, S., & Nishida, T. (2018). Language learning through conversation envisioning in virtual reality: A Sociocultural approach. In P. Taalas, J. Jalkanen, & S. Thouéšny (Eds.), *Future-Proof CALL: Language Learning as Exploration and Encounters-Short Papers from EUROCALL* (pp. 207-213). <http://doi.org/10.14705/rpnet.2018.26.838>
- Organisation for Economic Co-operation and Development (OECD). (2019). *Artificial Intelligence in Society*. OECD Publishing. <https://dx.doi.org/10.1787/eedfee77-en>
- Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, 29(3), 342-367.
- Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education*, 7(3), 135-153.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064-1082. <https://doi.org/10.1111/ajps.12103>
- Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010-1025.
- Shadiev, R., & Yang, M. (2020). Review of studies on technology-enhanced language learning and teaching. *Sustainability*, 12(2), 524. <https://doi.org/10.3390/su12020524>
- Song, P., & Wang, X. (2020). A Bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review*, 21(3), 473-486.
- Stockwell, G. (2007). Vocabulary on the move: Investigating an intelligent mobile phone-based vocabulary tutor. *Computer Assisted Language Learning*, 20(4), 365-383.
- Su, F., & Zou, D. (2020). Technology-enhanced collaborative language learning: theoretical foundations, technologies, and implications. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2020.1831545>
- Svensson, G. (2010). SSCI and its impact factors: A "Prisoner's dilemma"? *European Journal of Marketing*, 44(1/2), 23-33.
- Tran, B. X., Latkin, C. A., Vu, G. T., Nguyen, H. L. T., Nghiem, S., Tan, M. X., Lim, Z.-K., Ho, C. S. H., & Ho, R. (2019). The Current research landscape of the application of Artificial Intelligence in managing cerebrovascular and heart diseases: A Bibliometric and content analysis. *International Journal of Environmental Research and Public Health*, 16(15), 2699. <https://doi.org/10.3390/ijerph16152699>
- van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., Van der Ven, S., & Leseman, P. (2019). Social robots for language learning: A Review. *Review of Educational Research*, 89(2), 259-295.
- Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1), 79-105.
- Wang, C. P., Lan, Y. J., Tseng, W. T., Lin, Y. T. R., & Gupta, K. C. L. (2019). On the effects of 3D virtual worlds in language learning—A Meta-analysis. *Computer Assisted Language Learning*, 1-25.
- Wang, F., & Preininger, A. (2019). AI in health: State of the art, challenges, and future directions. *Yearbook of medical informatics*, 28(01), 16-26. <https://doi.org/10.1055/s-0039-1677908>
- Wijekumar, K. K., Meyer, B. J., & Lei, P. (2017). Web-based text structure strategy instruction improves seventh graders' content area reading comprehension. *Journal of Educational Psychology*, 109(6), 741-760.
- Yang, S. J. (2021). Guest Editorial: Precision education-A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 100025. <https://doi.org/10.1016/j.caeai.2021.100025>

- Zhang, R., & Zou, D. (2020). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, 1-47. <https://doi.org/10.1080/09588221.2020.1744666>
- Zou, D., Huang, Y., & Xie, H. (2019). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, 34(5-6), 751-777. <https://doi.org/10.1080/09588221.2019.1640745>
- Zou, D., Luo, S., Xie, H., & Hwang, G. J. (2020). A Systematic review of research on flipped language classrooms: theoretical foundations, learning activities, tools and research topics and findings. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2020.1839502>
- Zou, D., Xie, H., & Wang, F. L. (2018). Future trends and research issues of technology-enhanced language learning: A Technological perspective. *Knowledge Management & E-Learning: An International Journal*, 10(4), 426-440.

A Learning Analytics Framework Based on Human-Centered Artificial Intelligence for Identifying the Optimal Learning Strategy to Intervene in Learning Behavior

Fuzheng Zhao^{1,4}, Gi-Zen Liu², Juan Zhou³ and Chengjiu Yin^{1*}

¹Kobe University, Japan // ²National Cheng Kung University, Taiwan // ³Tokyo Institute of Technology, Japan //

⁴Jilin University, China // zhaofz635@gmail.com // gizen@mail.ncku.edu.tw // juan.z.kt@gmail.com //

yin@lion.kobe-u.ac.jp

*Corresponding author

ABSTRACT: Big data in education promotes access to the analysis of learning behavior, yielding many valuable analysis results. However, with obscure and insufficient guidelines commonly followed when applying the analysis results, it is difficult to translate information knowledge into actionable strategies for educational practices. This study aimed to solve this problem by utilizing the learning analytics (LA) framework. We proposed a learning analytics framework based on human-centered Artificial Intelligence (AI) and emphasized its analysis result application step, highlighting the function of this step to transform the analysis results into the most suitable application strategy. To this end, we first integrated evidence-driven education for precise AI analytics and application, which is one of the core ideas of human-centered AI (HAI), into the framework design for its analysis result application step. In addition, a cognitive load test was included in the design. Second, to verify the effectiveness of the proposed framework and application strategy, two independent experiments were carried out, while machine learning and statistical data analysis tools were used to analyze the emerging data. Finally, the results of the first experiment revealed a learning strategy that best matched the analysis results through the application step in the framework. Further, we conclude that students who applied the learning strategy achieved better learning results in the second experiment. Specifically, the second experimental results also show that there was no burden on cognitive load for the students who applied the learning strategy, in comparison with those who did not.

Keywords: Learning analytics framework, Analysis result application, Human-center AI, Learning strategy

1. Introduction

Learning analytics (LA) is viewed as a domain that combines data analytics and human judgment (Siemens, 2013). LA aims to reveal hidden patterns and generate actionable intelligence, which could provide timely intervention for students' learning behavior. A small number of efforts in LA have focused on the predictive analytics realm, in which techniques such as machine learning and deep learning were drawn upon (Xing & Du, 2018), and some analysis results have been applied to education for intervening learning behavior (Zhao et al., 2021). Some studies have shown that early prediction could help improve learning engagement (Gray & Perkins, 2019). In addition, systematic intervention strategies can successfully reduce the dropout rate (Choi et al., 2018).

While the predictive analytics domain has seen a surge, little is known about how to identify and apply the most suitable intervention strategy to students' learning behavior. Some concerns regarding intervention strategies after prediction analyses have been raised, as the application of prediction analysis results failed to show the expected efficacy. Bowers and Sprott (2012) found that some indicators may accurately predict which factors most affect academic performance but are unable to support effective interventions, owing to the lack of appropriate and effective interventions. Moreover, faced with the same analysis results of learning behavior patterns, the impact on students' learning behavior largely varies and is dependent on various intervention strategies (Rienties et al., 2016). As a result, a suggestion has been made that predictive analysis should go beyond simply predicting learning performance and should also inform instructors on appropriate intervention strategies (Barry & Reschly, 2012).

The application of analysis results is regarded as the final stage of the LA framework and is responsible for remedial actions for students (Clow, 2012). In previous research, it was found that the application step in the current LA frameworks (Campbell et al., 2007; Chatti et al., 2012; Dron & Anderson, 2009; Elias, 2011; Siemens, 2013), including the ReCoLBA framework that we proposed (Zhao et al., 2021), lacks guidance on how to transform the analysis results into corresponding implementable strategies, especially learning strategies.

Moreover, human-centered Artificial Intelligence (HAI) has been emerging as a new development trend. It not only seeks to consider the human condition when designing the AI but also identifies human learning patterns

and facilitates timely intervention by artificial intelligence (AI) techniques and big data (Tsai et al., 2020). In particular, it requires the computation and application of AI algorithms with machine intelligence to be trustworthy and responsible, thereby enhancing the welfare of humankind. To increase the trustworthiness of analysis results produced by AI algorithms, make them responsible for computation, and consider human welfare when applying them, we proposed an LA framework that considers HAI, emphasizing the result application step.

2. Literature review

2.1. Opportunities of HAI for the LA framework

The AI research goals in relation to education consist of prediction, structure discovery, and relationship mining (Yin & Hwang, 2018). Considering the new development trend of AI and its typical application, machine learning, Shneiderman (2020) envisions the use of HAI as inevitably on the rise and taking an evolutionary direction, considering human conditions and contexts in its design and application. As an important branch of HAI, the essence of LA is to apply big data and AI to identify at-risk students and intervene promptly. In particular, HAI requires explainable, trustworthy, and responsible computation for AI algorithms and applications with machine intelligence, thereby improving human welfare (Yang et al., 2021).

2.2. Problems in LA frameworks

Although studies of LA frameworks have been undertaken, they have not examined the details of the implementation of the analysis result application in those frameworks. Campbell et al. (2007) initially proposed the “Five Steps of Analytics” framework, comprising the steps “capture,” “report,” “predict,” “act,” and “refine.” The term “act” refers to the application of analysis results. As a type of intervention, they suggest supporting the above applications with a personal phone call or e-mail.

As the introduction of LA to various learning environments increased, different perspective-based explorations on the framework also saw urgent development. Dron and Anderson (2009) defined their “Collective Application Model” framework taking into consideration the characteristics of e-learning. As it highlighted information gathering, processing, and presentation, the analysis result application step was excluded from the framework, which comprised only “select,” “capture,” “aggregate,” “process,” and “display.” After a comparison of existing frameworks, Elias (2011) proposed a comprehensive framework comprising “select,” “capture,” “aggregate and report,” “predict,” “use,” “refine,” and “share.” Regarding the application step, it only provides a brief description of attempts to improve the learning system.

The Chatti framework (Chatti et al., 2012) incorporates analysis and its corresponding application into one step to compose the processing step, with the remaining two steps pre-processing and post-processing. In contrast, Siemens (2013) not only added an application step to his framework but also highlighted the purposes of the step, such as “intervention,” “optimization,” “alters,” “warning,” “guiding” / “nudging,” “systemic,” and “improvement.” Two issues exist in these frameworks. First, the reason for the analysis results cannot be sufficiently understood due to the lack of confirmation of the analysis results by AI algorithms, reducing the explanation and trustworthiness of analysis results. Second, existing frameworks cannot provide specific ways to apply the analysis results of AI algorithms, as shown in Table 1. In this context, analysis results cannot be successfully translated into the proven application strategy and thus cannot effectively support human welfare. The former was addressed in the ReCoLBA framework (Zhao et al., 2021), and the latter is the focus of this study.

Table 1. Features and drawbacks of the application step in the existing frameworks

Title	Features	Drawbacks
“Five Steps of Analytics” framework	Specific intervention approaches (phone calls or e-mail)	Limited application range
“Collective Application Model” framework	Skips the application step	Lack of function
Elias’ framework	Only defines the application	No specific guidance steps
Chatti’s framework	Considers the application a sub-step of data processing; no other details are available	The independent properties of the application are not established
Siemens’ framework	Describes the purposes of the application	Lack of application methods
ReCoLBA framework	Describes the stakeholders of the application	Lack of specific guidance

2.3. Low application effect in analysis result application

The application of analysis results in LA is a key step, aiming to apply the analysis results to educational practice for monitoring, prediction, intervention, assessment, adaptation, personalization, recommendation, and reflection (Chatti et al., 2012). To achieve the intended application targets, it is critical to fit the analysis results into implementations of education activities when developing an application.

However, having good analysis results is not always successful in facilitating educational activities, and it is especially crucial to apply them correctly (Viberg et al., 2018). Hanna (2004) attributes this to the ambiguity of the analysis results: the results may yield useful insights and clues without providing definitive answers. Furthermore, Saks et al. (2018) found that the ambiguity of the analysis results has caused most users to hold an ambiguous view of application effectiveness. On the other hand, most studies have concentrated on how to meet application requirements along with the advent of these learning scenarios, tools, and data. In contrast, little attention has been given to concrete steps to use the analysis results to create an optimal application strategy.

2.4. Big data-driven education application strategy development

The term evidence-driven education (EDE) was first used by Hargreaves (1997) and was inspired by evidence-driven practice in the field of medicine. EDE aims to bridge the research-to-practice gap in teaching as well as to shift the driving force of instructional programs and practices to evidence rather than ideology, faddism, politics, and marketing (Davies, 1999). EDE is supported by the findings of multiple, high-quality, experimental studies (Cook et al., 2008) and quantitative analysis (Moran & Malott, 2004), which provides sound evidence that an educational practice truly works. Kuromiya et al. (2020) used evidence obtained from reading behavior analysis conducted when students were learning via a learning management system to evaluate whether a new learning intervention has a positive learning effect.

In summary, significant scope for exploration remains in the application of analysis results, especially in the context of current LA frameworks. To make up for the drawbacks of the application step in all frameworks, an LA framework based on HAI, also called the HAILA framework, was designed to identify the optimal learning strategy and provide an application-step sequence.

The research questions (RQs) to be addressed in this study are as follows:

- RQ1. Is the proposed framework conducive to effectively identifying the optimal learning strategy? If so, how?
- RQ2. What is the effect of the learning strategy identified by the proposed framework?

3. Research on the HAILA framework

The HAILA framework is based on but differs from the ReCoLBA framework, as shown in Table 2. This framework also contains steps related to data collection, data processing, data analysis, result confirmation, and result application, as shown in Figure 1. More importantly, the HAI concept is introduced for the design process of this framework, with a focus on refining the application step to identify the optimal learning strategy to intervene in learning behavior. This framework aims to increase the dependability of AI algorithms' analysis and results, ensure the accountability of the application of their analysis results, intervene in educational practice in a timely and accurate manner, and improve learning outcomes.

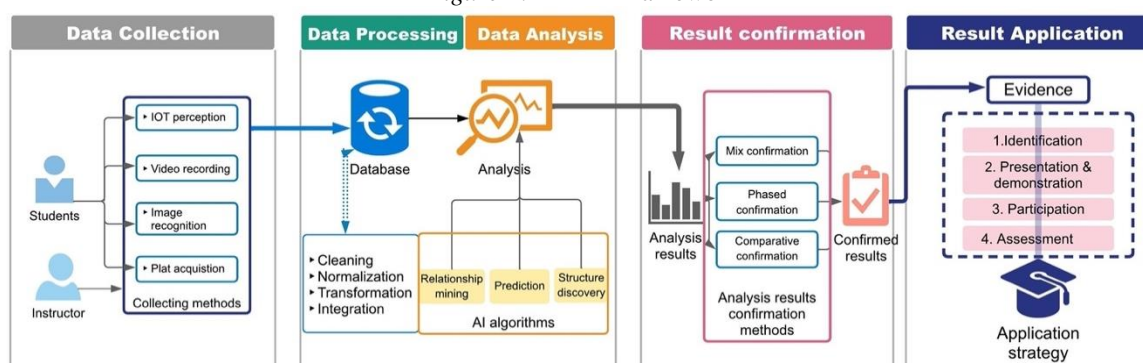
Table 2. Differences and similarities between ReCoLBA and HAILA frameworks

Name	Differences	
ReCoLBA	Purpose	Confirming analysis results produced by AI algorithms
	Design concept	Does not consider HAI-driven design
	Disadvantage	Lacking specific guidance in the application-step sequence
HAILA	Purpose	Confirming analysis results by AI algorithms and finding the application strategy
	Design concept	Considers HAI-driven design
	Advantage	Has a specific application-step sequence

Diverging from other frameworks that do not consider HAI, the HAILA framework not only includes a result confirmation step to identify the accuracy of analysis results produced by AI algorithms but also provides an application-step sequence to transform the complex, abstract results obtained from AI algorithms into an

application strategy. Thus, instructors' understanding of AI algorithms' analysis results will increase (Zhao et al., 2021), and the effectiveness of AI algorithm analysis applications is expected to be guaranteed.

Figure 1. HAILA framework



3.1. Data collection

In the LA field, data collection refers to a process in which information is gathered from various educational environments using a variety of techniques, such as video recording, image recognition, platform acquisition, and IOT perception. To provide a sound data foundation for successive steps, data collection should be characterized by timeliness, consistency, accessibility and convenience, accuracy, and responsiveness (Russell & Taylor, 2008).

3.2. Data processing

This consists of imputation of missing values, data noise identification, data integration, data cleaning, data normalization, and data transformation. This step aims to offer the most suitable data for analysis through the steps of retrieving, identifying, manipulating, modifying, and replacing.

3.3. Data analysis

The analysis goals include prediction, structure discovery, and relationship mining. To accomplish these goals and consider the characteristics of various education scenarios, a total of 12 methods under three categories corresponding to the above goals are applied in practice. Four methods, namely association rule mining, correlation mining, sequential pattern mining, and causal data mining, support the relationship mining goal. The structure discovery goal can be achieved using clustering, factor analysis, knowledge inference, and network analysis. The prediction goal primarily depends on classification, regression, latent knowledge, and estimation.

3.4. Result confirmation

Confirmation of the analysis results is performed before application. The reasons for the analysis results can be determined based on the confirmation methods included in the framework design. The confirmation step comprises mixed, phased, and comparative confirmation.

3.5. Result application

As part of the curriculum design, learning strategies are primarily responsible for the realization and completion of the instruction objectives. To find the optimal application strategy, an evidence-driven education policy was employed in the application-step sequence, which is responsible for building a corresponding link between the identified learning strategy and the analysis result. The evidence-driven education policy establishes the hypothesis of matching between the analysis results and the application strategy. Hence, the application-step sequence helps instructors determine the optimal application strategy. In addition, the evidence for constructing hypotheses using the optimal applied strategy is based on the analysis of data from any learning scenario and is not for one specific such scenario. Therefore, the application strategy obtained can be applied to any learning

scenario. Four steps were designed as part of the application-step sequence, including identification, presentation and demonstration, participation, and assessment. The hypothesis regarding the analysis results and application strategies is fulfilled in the identification step, which reflects the design concept of evidence-based education.

3.5.1. Identification

As the existing frameworks lack specific application guidance, they do not provide the function of strategy identification. This step aims to determine an optimal learning strategy suitable for the confirmed analysis results. Thus, the degree to which the identified learning strategy matches the confirmed analysis results determines the application effect.

3.5.2. Presentation and demonstration

This step allows the instructor to disseminate information to learners using verbal information in writing, and visual symbols. It aims to gain learners' attention, inform learners of objectives, and combine new strategies with prior knowledge. When illustrating, concise and concrete steps for implementing a strategy are crucial.

3.5.3. Participation

This application step allows the learner to use the identified strategy to affect learning achievement. In this step, learners are asked to participate in a learning scenario and retry the strategy until they can use it successfully in real-world practice (Dick et al., 2015).

3.5.4. Assessment

This step has two dimensions. The first is to provide accurate feedback regarding learner performance when using this strategy, focusing on academic achievement in regard to the learning content (Dick et al., 2015). The second is to measure the cognitive load on the learner, which can reveal the impact of the learning strategy. The combination of academic performance and cognitive load measures is considered to provide a reliable estimate of the efficiency of instruction methods (Paas et al., 2003).

4. Experiment design

To examine the effectiveness and contribution of the HAILA framework, a case study was conducted using two independent experiments. The first experiment is designed to analyze reading behavior data and to confirm the trustworthiness of its analysis results. The learning behavior that contributes most to learning achievement in the experiment is expected to be determined and will be used as evidence for an application strategy to intervene in learning behavior. To verify the effectiveness of SQ3R, which is hypothesized based on the evidence that SQ3R could encourage students to turn back to re-read pages, another experiment is conducted.

4.1. Experiment for confirming framework effectiveness

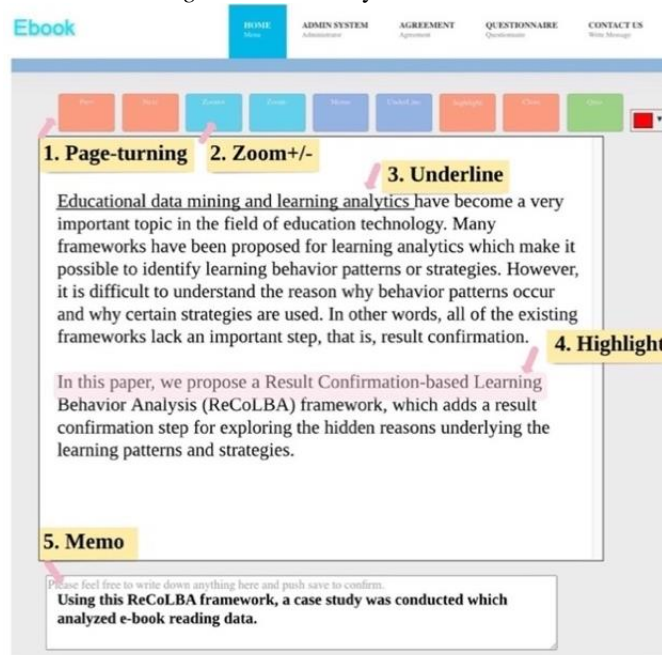
A total of 234 freshmen were recruited, among which the gender distribution was 65 males and 169 females, with an average age of 19 years old. Five experimental steps were employed according to the HAILA framework, namely (1) collecting students' reading behaviors using an e-book system, (2) using a machine learning library (scikit-learn) to process raw data, (3) utilizing classification prediction algorithms and feature importance calculation methods in the analysis step, (4) adopting different algorithms to confirm the analysis results, and (5) applying an identified learning strategy.

4.1.1. Data collection by e-book system

An e-book system was developed (Yin et al., 2018) to capture learners' reading learning behavior. Learners can conveniently read materials using several operating tools, such as (1) turning the page forward or backward, (2)

resizing the view by zoom, (3) creating a memo, (4) adding or removing underlining in the learning material, and (5) adding or deleting highlights with a variety of color options, as shown in Figure 2.

Figure 2. E-book system interface



All the learning behavior observed in the learning activities was recorded in a data log with 12 data features, as shown in Table 3. Among the e-book features, the most used were the tools associated with basic reading behaviors, such as Next (NE), Prev (PR), Highlight (HL), Underline (UL), Memo (ME), Bookmaker (BM), Read time (RT), and Read page (RP). In addition, the BacktrackRate (BR) feature is a statistical ratio dividing Next and Prev, which provides a new view on repeated learning behavior. The equipment used for learning, such as a PC, mobile phone (Mobile), and tablet, was adopted as a parameter.

Table 3. Samples of reading behavior data derived from the e-book system

Id	PC	Mobile	Tablet	BR	ME	HL	UL	PR	NE	RT	RP	BR
1	0	1	0	7	1	29	3	9	12	60	51051	0.75
2	0	0	1	8	0	20	4	79	96	107	42043	0.823

4.1.2 Data processing using scikit-learn

The data processing steps were divided into two categories. The first consisted of basic processing, mainly involving the removal of missing values to maximize the validity of the data and standardization to unify the values of each data feature. As a result, valid data from 229 participants remained, and all data values were compressed from 0 to 1. The second category focused on the analysis method-specific data preparation. Considering the demands of data balance for binary classification prediction (Krawczyk, 2016), we adjusted the passing score of 60, which had a huge gap in proportion, to 70. In addition, splitting data into train- and test-datasets is necessary as part of regular data processing. The *train_test_split* function built into scikit-learn was adopted to achieve the specific splitting ratio of 3:7, in which the training dataset accounted for 70%.

4.1.3. Data analysis by decision tree model

In this study, the decision tree model was used to explore which data feature contributes most to predicting academic performance (Hamoud et al., 2018; Mesarić & Šebalj, 2016). After the prediction model was created and tuned by scikit-learn, we obtained the final model with five metrics: accuracy (0.739), F1-score (0.763), recall (0.763), precision (0.763), and AUC (0.81). Following the rule of thumb, the predictive performance of this model can be viewed as fair. Note that an AUC score over 0.8 represents good comprehensive performance in terms of a model's effectiveness.

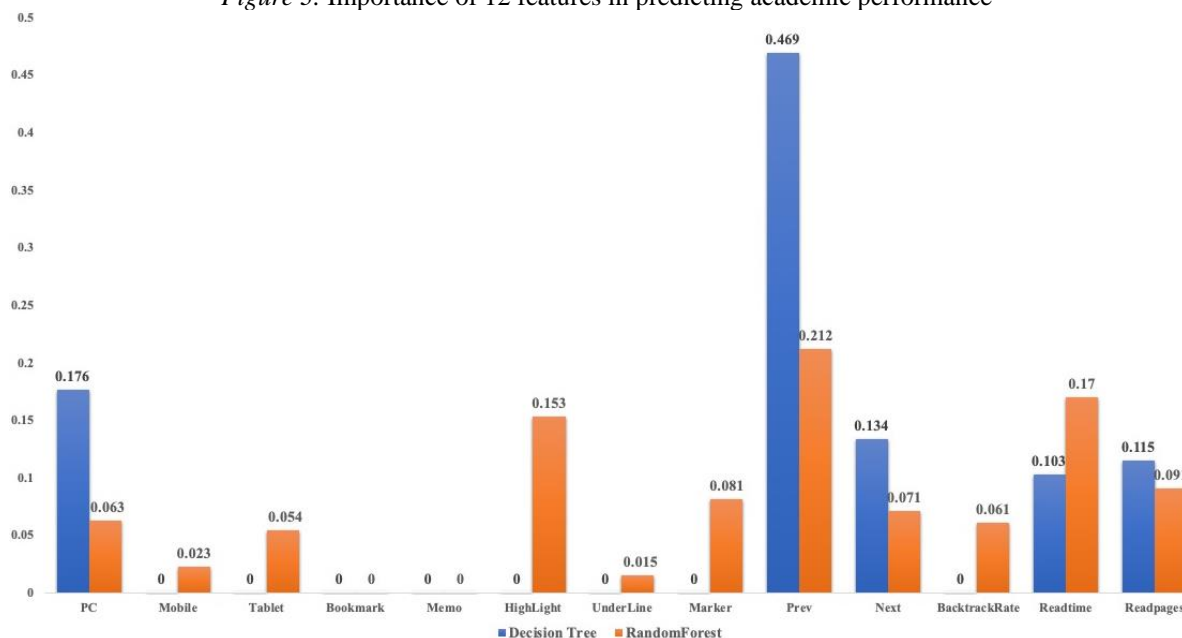
Scikit-learn offers an impurity-based feature importance calculation function oriented to the decision tree model. Subsequently, the 12 data features were analyzed using this function. As fundamental splitting parameters to calculate the feature importance for classification prediction, the Gini Index and Information Gain prevail in terms of splitting criteria. Raileanu and Stoffel (2004) found that there are no obvious differences by comparing the efficiency of splitting features for tree models. Moreover, the Gini Index has proved to be better than the other splitting parameters specifically for unbalanced datasets (Park & Kwon, 2011). As our sample size of students who fail the exam is unbalanced relative to those who passed, accounting for a minority, the Gini Index was adopted to calculate the probability weighting of each node in the tree model, with values ranging between 0 and 1, where 0 expresses the purity of classification. The values for PC, Prev, Nex, Read time, and BacktrackRate are 0.176, 0.496, 0.134, 0.103, and 0.115, respectively, and the other features have values of 0. It was found that the Prev feature had the greatest impact on prediction performance. In other words, students who have Prev learning behavior are more likely to pass the exam and obtain better academic performance.

4.1.4. Result confirmation by a comparative method

The most effective data feature, Prev, was successfully identified for predicting students' academic performance. However, it was unclear whether this analysis result was sufficiently accurate to obtain the same results in other algorithms. Guided by the analysis result confirmation step, a comparative confirmation method was employed to confirm the correctness and reproducibility of the analysis results. As an innovative algorithm based on the tree model (Liaw & Wiener, 2002), the random forest algorithm outperforms the other algorithms (Breiman, 2001) and is commonly used for binary prediction models.

In the confirmation step, the same experimental conditions were set, and only the selected algorithm was shifted from the decision tree to the random forest model. Following the rule of thumb, the prediction based on the random forest model was also acceptable in terms of accuracy (0.753), precision (0.729), recall (0.794), F1-score (0.76), and AUC (0.75) according to the results of data feature importance in the decision tree and random forest models, as shown in Figure 3. This shows that the Prev parameter consistently contributes the most in the prediction of academic performance, indicating that students who show the Prev learning behavior have a higher probability of passing the exam.

Figure 3. Importance of 12 features in predicting academic performance

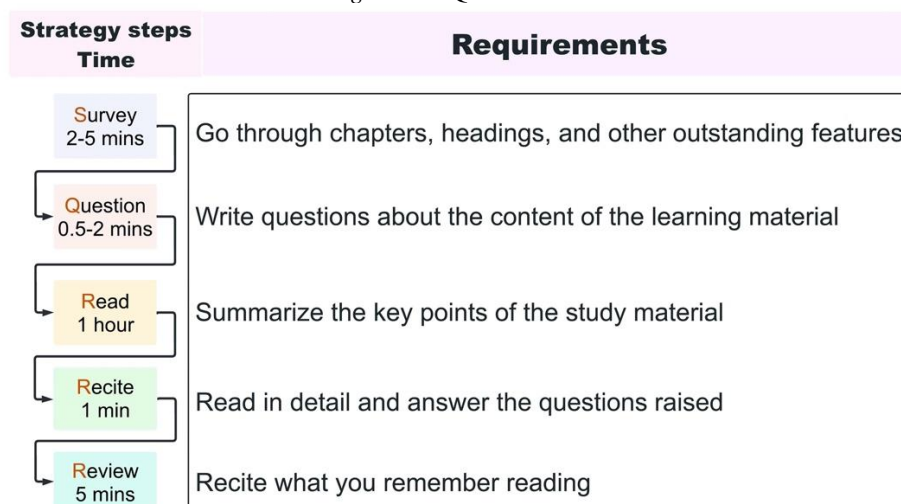


4.1.5. Result application by SQ3R

After the researchers analyzed learning behavior using AI algorithms, the learning behavior that contributed most to learning efficiency is to be considered evidence of strategy making. Moreover, a hypothesis regarding application strategy was presented to facilitate the improvement of learning efficiency. First, the identification step was entered. The Prev behavior raised the need to increase learning frequency. Consistent with the demand

in the learning frequency, we found that the SQ3R learning strategy was primarily designed for university students reading academic textbooks (Li et al., 2014). It is a comprehensive reading method comprising five steps: survey, question, read, recite, and review (Flippo & Bean, 2018). Importantly, multiple SQ3R learning steps can increase the occurrence of repeated learning behaviors (Huber, 2004). As shown in Figure 4, short-term memory achieved by repeatedly turning pages is necessary when students are quickly browsing notable features and writing down reading questions in survey and question steps. In subsequent steps, reciting and confirming what has been read are also realized by turning pages. To this end, it is hypothesized that SQ3R facilitates the occurrence of the Prev behavior.

Figure 4. SQ3R schedule



In the presentation and demonstration step, an introduction of SQ3R explaining how to use it in the e-book system was provided by an instructor. In Figure 4, a schedule was also designed based on previous instruction experience. For example, the survey step is recommended to last 2 to 5 min, including reviewing notable features in the textbook by turning pages and highlighting or underlining keywords, followed by 30 seconds to 2 mins to write down questions as memos, 1 hour for reading in detail, 1 min for reciting the memory, and 5 mins for summarizing and reviewing the material that was the emphasis of the learning experience by quickly turning pages. Subsequently, learners could undertake the participation step with the guidance of the presentation and demonstration steps. Finally, a post-test regarding the learning content and a test of cognitive load provide feedback to learners in the assessment step.

4.2. Experiment for verifying the framework contribution

This study examines the contribution of this framework by exploring the effectiveness of the SQ3R. Particularly, this experiment aims to evaluate the effectiveness of the SQ3R in helping students improve their academic performance. To this end, an e-book system was used to complete the above experiment to evaluate whether significant changes in academic performance and reading behaviors were related to the SQ3R application.

4.2.1. Participants

Thirty-seven male and 32 female freshmen, aged from 19 to 21 years, participated in this study. All participants were randomly assigned to two groups: the experimental group and the control group. The 35 participants in the experimental group were asked to read a learning material using the SQ3R, while the control group, which comprised 34 participants, was assigned to read the learning material without the guidance of the SQ3R.

4.2.2. Measuring method

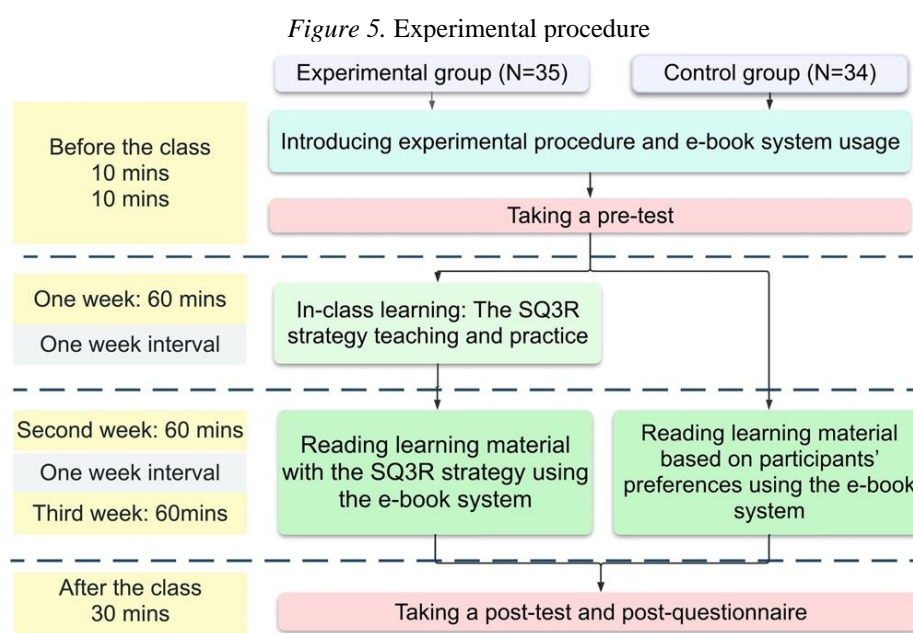
The measuring method comprised a pretest and a post-test. Before the experiment, a pretest was conducted to assess whether the two groups had the same equivalent prior knowledge regarding the learning content and the SQ3R. This pretest consisted of 10 multiple-choice items. The post-test aimed to measure whether the SQ3R was beneficial to participants' academic performance. This test was similar to the pre-test, comprising 10 multiple-

choice items related to the learning emphasis in the upcoming learning materials. The pretest and post-test were both scored on a 10-point scale.

A post-questionnaire was employed to identify participants' cognitive load when using the SQ3R, and their attitude about introducing new learning technologies into the reading environment. Based on the measurement created by Sweller (1988), a questionnaire to investigate cognitive load was modified. The developed questionnaire consisted of eight questions with two dimensions: mental load and mental effects. Each question was scored on a 5-point scale, where 5 represented "strongly agree" and 1 indicated "strongly disagree." The Cronbach's alpha values of the two dimensions were 0.91 and 0.95.

4.2.3. Experimental procedure

According to Figure 5, this study consisted of a pretest, a reading-based learning activity using the e-book system, a post-test, and a post-questionnaire on cognitive load and learning strategy acceptance, which all together took 2 weeks.



Initially, an orientation was provided for participants to introduce the experimental procedure and the e-book system operation and precautions. Following the orientation, the experimental and control groups were asked to complete a pretest to evaluate their knowledge of the SQ3R. Afterward, in-class instruction of the SQ3R was provided only to the experimental group over 60 mins during the first class, followed by an independent practice targeting the SQ3R proficiency during 1 week. Then, the experiment proceeded to two learning activities that took 60 mins each, with an interval of 1 week in between. The participants in the experimental group read the learning material using the SQ3R, whereas those in the control group read the same learning material, but the reading strategy was based on their preferences. Subsequently, a post-test and post-questionnaire, which took 30 mins, were conducted with the two groups.

5. Experimental results

We first explored the learning behavior that contributes most to learning achievement by analyzing learning behavior in an e-book system and used the analysis results as evidence for developing application strategies, hypothesizing that the developed strategies could promote learning achievement. Finally, the hypothesis has been verified by the following experiment. Hypothesis testing utilizing 61 valid samples, including an independent sample *t*-test and one-way analysis of variance (ANOVA) in SPSS Statistics, was employed to examine the effectiveness of the SQ3R in improving academic performance and affecting learning behavior.

The experimental results show that the experimental group using the SQ3R significantly outperformed the control group in terms of academic performance, even with equivalent prior knowledge. Furthermore, regarding

Prev, Read time, and Read page behaviors, the experimental group showed a significantly higher number of occurrences. In addition, the above differences are also verified by data visualization. Finally, the cognitive load test revealed that the SQ3R did not impose additional cognitive load on students' learning.

5.1. Analysis of academic performance

The pretest results showed that the standard deviation and mean values were 2.18 and 6.67 for the experimental group and 2.15 and 7.35 for the control group. According to the *t*-test results ($t = -1.23$, $p > .05$; Table 4), no statistically significant differences were found between the two independent groups. Thus, all participants in both groups were known to have equivalent prior knowledge regarding the SQ3R.

After the learning activity, ANOVA was conducted to determine whether there was any statistically significant difference in the post-test results between the two groups. Using the groups as a fixed factor and post-test scores as the dependent variable, a common assessment for homogeneity of variance was performed using Levene's test. A Levene's test score above 0.05 ($F = .09$, $p = .75$) indicated that this test is robust to violations of the assumption. It was concluded that the experimental group was statistically different from the other group. Based on the statistical results ($F = 32.86$, $p < .01$) shown in Table 5, the participants who learned with the SQ3R showed significantly better academic performance than those who did not. In other words, the SQ3R was helpful for participants in improving their academic performance.

Table 4. Descriptive data and *t*-test result of the pretest results

Variable	Group	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Pretest	Control group	30	6.67	2.18	-1.23*
	Experiment group	31	7.35	2.15	

Note. * $p > .05$.

Table 5. Descriptive data and ANOVA result of the post-test results

Variable	Group	<i>N</i>	Mean	<i>SD</i>	<i>F</i>
Post-test	Control group	30	4.50	1.77	32.86**
	Experiment group	31	7.06	1.71	

Note. ** $p < .01$.

5.2. Analysis of learning behavior

First, a *t*-test was used to analyze the repeated learning behavior based on Prev, Read time, and Read page in the two groups, as shown in Table 6. Based on the hypothesis that the SQ3R contributes to the emergence of repetitive learning behavior such as Prev, the SQ3R was applied in the experiment particularly to facilitate the occurrence of Prev behavior and achieve this purpose. For Prev, the mean and standard deviation were 24.43 and 26.07 for the control group and 60.39 and 35.71 for the experimental group. Moreover, the *t*-test result ($t = -4.50$, $p < .01$) showed that there was a significant difference between the two groups, implying that the SQ3R was able to promote the occurrence of Prev.

Analyzing the Read time variable alone, the difference between the two groups is apparent, and the *t*-test result ($t = -2.35$, $p < .05$) further verifies this conclusion, as shown in Table 6. However, the Read time variable was combined with the Read page variable for analysis. It is worth noting that the participants using the SQ3R took more time than those who did not, but after accounting for the Read page variable based on the *t*-test result ($t = -4.83$, $p < .01$), it was found that the former read nearly twice as efficiently as the latter. On average, the experimental group read approximately 3.01 pages per minute, compared with the control group's 1.59 pages per minute.

Second, we used the *t*-test to explore whether the SQ3R affects other reading behaviors. The results showed that there was a significant difference between the two groups in terms of Next ($t = -3.42$, $p < .01$), Memo ($t = -3.16$, $p < .01$), Underline ($t = -3.62$, $p < .01$), and Bookmark ($t = -4.97$, $p < .01$), though not for Highlight ($t = .06$, $p > .05$). These findings reveal that the SQ3R also promoted the occurrence of reading behaviors such as Next, Memo, Underline, and Bookmark. Regarding Highlight, no significant difference was found between the two groups.

Table 6. Descriptive data and *t*-test results for learning behavior

Variable	Group	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Prev	Control group	30	24.43	26.07	-4.50**
	Experiment group	31	60.39	35.71	
Read time	Control group	30	0:49:18	0:20:31	-2.35*
	Experiment group	31	1:00:30	0:16:30	
Read page	Control group	30	78.57	61.30	-4.83**
	Experiment group	31	182.32	101.96	
Underline	Control group	30	5.40	10.19	-3.62**
	Experiment group	31	37.52	48.16	
Highlight	Control group	30	5.30	18.22	-.060
	Experiment group	31	5.52	8.57	
Bookmark	Control group	30	.43	2.37	-4.97**
	Experiment group	31	6.90	6.82	
Memo	Control group	30	.00	.000	-3.16**
	Experiment group	31	2.19	3.85	
Next	Control group	30	43.00	28.42	-3.42**
	Experiment group	31	72.00	37.04	

Note. * $p < .05$; ** $p < .01$.

Third, Figures 6 and 7 show learning patterns in terms of the time distribution of reading behaviors. The X-axis represents the time participants spent reading the material in the e-book system, and the Y-axis indicates the reading behaviors. The upper part of the two figures is the time distribution for each SQ3R step, where the time division between steps is based on the application guidance designed in the framework application step. Although there exists an obvious time division between various steps, this does not mean that all learning behaviors have a similar distribution to the preset learning steps.

Figure 6. Distribution of reading behaviors of the experimental group using the e-book system



Figure 7. Distribution of reading behaviors of the control group using the e-book system

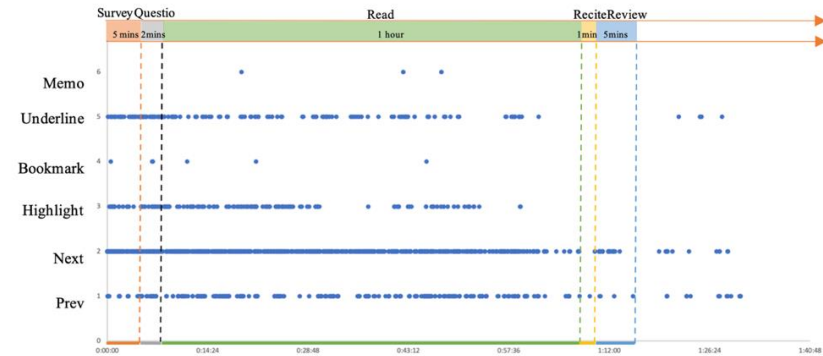


Figure 6 shows a rough distribution trend of reading behaviors representing three stages in terms of period: 0 min to 3 min, 3 min to 1 h 12 min, and 1 h 12 min to the end. Combining the five SQ3R divisions shows that in Stage 1, Prev, Underline, and Bookmark appear intensively in a short period, which indicates that the participants repeatedly read the material using the survey step, marking some knowledge points in a short time. Stage 2

accounts for most of the reading behavior. This stage might include the question and reading. The reading behaviors in Stage 3 primarily comprise Prev and Next; the other behaviors account for a smaller proportion. The time distribution of reading behaviors in the experimental group is consistent with the SQ3R, which suggests that participants spent 2 to 3 min on the first survey step, 2 min on the question step, and 5 min each on the reciting and review steps.

Figure 7 shows that the control group is divided into two stages, in which Prev and Next occur continuously and are the most frequent. Memo and Bookmark are less frequent, scattered, and irregular. In the first stage, Highlight and Underline appear frequently. Highlight has two distinct distributions, and it appears intensively and continuously for 30 min. Underline remains stable across several occurrences.

5.3. Analysis of cognitive load

A cognitive load post-questionnaire, which includes two test dimensions, mental load and mental effort, was employed to investigate the differences between the two groups, in terms of learning pressure and load on the participants. Because the mental load effect depends on the information being processed, which imposes a heavy cognitive load (Sweller, 1988), the first dimension focuses on the pressure caused by the amount of information the participants process. In addition, the second dimension carries out the mental effort test, which reflects the controlled consumption of psychological information processing resources in the cognitive process (Sweller, 1988; Hwang & Chang, 2011).

The *t*-test results in Table 7 show that for the mental load dimension, the mean and standard deviation were 11.91 and 3.25 for the control group and 11.17 and 2.35 for the experimental group. No significant difference was found between the two groups ($t = 1.01, p > .05$), implying that the SQ3R did not increase the pressure of information amount on participants. Regarding mental effort, no significant difference was found between the two groups ($t = 1.66, p > .05$). Therefore, the SQ3R did not exert additional pressure on participants in terms of mental load or mental effort. All participants in both groups had a moderate level of learning pressure, as indicated by the similar standard deviation concentrating at 3 for both groups.

Table 7. Descriptive data and *t*-test results for cognitive load

Variable	Group	<i>N</i>	Mean	<i>SD</i>	<i>t</i>
Mental load	Control group	30	11.91	3.25	1.01*
	Experiment group	31	11.17	2.35	
Mental effort	Control group	30	10.47	3.27	1.66*
	Experiment group	31	9.16	2.82	

Note. * $p > .05$.

6. Discussion

This study explored whether the HAILA framework would affect the identification of optimal learning strategies. RQ1 investigated the effectiveness of the evidence-driven framework to offer guidance for transforming analysis results into the most suitable application strategy. For the application step performance, the SQ3R was obtained and optimally matched with the analysis results of the first experiment. The need for analysis result application has been proven by previous studies (Barry & Reschly, 2012); however, no previous study has provided concrete guidance for implementing the application. Inspired by HAI, evidence-driven education was incorporated into the design of the application step. Thus, this study explored the extent to which evidence-based education can facilitate the identification and transformation of analysis results into the optimal strategy. The experiment results show that evidence-driven education can sufficiently support application step design.

There remain downsides to the use of AI in LA, particularly related to algorithmic bias (Carter & Egliston, 2021). For example, decision-making based on AI analytics with unrepresentative datasets and algorithm design bias results in not only incomprehensible analysis results but also inappropriate or inapplicable result applications. Specifically, training AI algorithms on historical and complicated learning behavior data may reinforce the difficulty of understanding how it potentially undermines the learning behavior patterns. In that case, the HAILA framework contributes to increasing the interpretability of learning behavior patterns, not simply for exploring learning behavior patterns themselves. For example, the database on which AI algorithms are based is inevitably biased in terms of gender, family background, and ethnicity, resulting in the data itself containing bias (West et al., 2018). Moreover, concerns exist that AI analysis results can trigger artificial biases

against learning behavior without an inappropriate or inapplicable application strategy, which reduces the trustworthiness of the AI analytics application. Therefore, the proposed framework with HAI consideration focuses on identifying the best application strategy that matches the analysis results. This function helps avoid artificial biases on learning behavior caused by inappropriate application strategies.

7. Conclusions

The HAILA framework is significantly effective in terms of analysis result application, and it highlights an implementable way to identify and apply the optimal application strategy, particularly concerning learning strategies. To verify the effectiveness of the modified framework and application strategy, two independent experiments were conducted. The results of the first experiment show that a learning strategy that best matched the analysis results was found through the application step in the framework. In addition, the findings of comparative experiments showed that students who applied the learning strategy achieved better learning results. This result is consistent with Li's et al. (2014) study showing that the SQ3R contributes to good learning achievement. However, it is unclear whether the SQ3R demands additional cognitive load. Moreover, a *t*-test showed that the experimental-group students who applied the learning strategy were not burdened with additional cognitive load, compared with the control-group students.

In contrast to the current analysis result application approaches, the HAILA framework consists of four steps, in which there are two design focuses. In particular, evidence-driven education is used to determine the optimal learning strategy, and the cognitive load test provides feedback on the application of the learning strategy. According to the results of the experiment, it was concluded that the SQ3R can improve academic performance by motivating the occurrence of repeated learning. Moreover, statistical results showed that the SQ3R helps increase the frequency of Prev learning behavior. In terms of the cognitive load test, there was no significant difference between the students who used the SQ3R and those who did not.

One of the purposes of HAI is to accurately identify at-risk students using AI algorithms and provide timely intervention that considers human education welfare. Some students are inevitably at risk of low academic performance; therefore, how to intervene promptly is a crucial problem. Based on previous studies, most research emphasizes identifying students who are at risk in terms of academic performance by analyzing learning behavior rather than applying learning strategies to overcome these risks. This study's primary contribution is that it succeeded in not only enhancing the trustworthiness of AI algorithms analysis results and verifying which factors contribute most to learning performance but also determining the optimal learning strategy for intervention in learning behavior to guarantee the effectiveness of AI algorithm analysis application.

Acknowledgement

This research was partially supported by the Grants-in-Aid for Scientific Research Nos. 21H00905 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

References

- Barry, M., & Reschly, A. L. (2012). Longitudinal predictors of high school completion. *School Psychology Quarterly*, 27(2), 74–84. <https://doi.org/10.1037/a0029189>
- Bowers, A. J., & Spratt, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A Growth mixture model analysis. *The Journal of Educational Research*, 105(3), 176–195. <https://doi.org/10.1080/00220671.2011.552075>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A New tool for a new era. *EDUCAUSE Review*, 42(4), 40–42.
- Carter, M., & Egliston, B. (2021). What are the risks of virtual reality data? Learning analytics, algorithmic bias and a fantasy of perfect data. *New Media & Society*, 21(10), 1–20. <https://doi.org/10.1177/14614448211012794>
- Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A Reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5/6), 318–331. <https://doi.org/10.1504/ijtel.2012.051815>

- Choi, S. P. M., Lam, S. S., Li, K. C., & Wong, B. T. M. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Educational Technology & Society*, 21(2), 273–290.
- Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In B. Wellman (Ed.), *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge: SESSION. Institutional perspectives* (pp. 134–138). ACM. <https://doi.org/10.1145/2330601.2330636>
- Cook, B. G., Tankersley, M., Cook, L., & Landrum, T. J. (2008). Evidence-based practices in special education: Some practical considerations. *Intervention in School and Clinic*, 44(2), 69–75. <https://doi.org/10.1177/1053451208321452>
- Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, 47(2), 108–121. <https://doi.org/10.1111/1467-8527.00106>
- Dick, W., Carey, L., & Carey, J. O. (2015). *The Systematic design of instruction* (8th ed.). Pearson.
- Dron, J., & Anderson, T. (2009). On the design of collective applications. In B. Randall (Ed.), *2009 International Conference on Computational Science and Engineering: Vol.4. Social intelligence and networking VII* (pp. 368–374). IEEE. <https://doi.org/10.1109/CSE.2009.469>
- Elias, T. (2011). *Learning analytics: Definitions, processes and potential* (Computer Science; Corpus ID: 16906479) [Data set]. Semantic Scholar.
- Flippo, R. F., & Bean, T. W. (Eds.). (2018). *Handbook of college reading and study strategy research* (3rd ed.). Madison Taylor & Francis Routledge. <https://doi.org/10.4324/9781315629810>
- Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, 131, 22–32. <https://doi.org/10.1016/j.compedu.2018.12.006>
- Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 26–31. <https://doi.org/10.9781/ijimai.2018.02.004>
- Hanna, M. (2004). Data mining in the e-learning domain. *Campus-Wide Information Systems*, 21(1), 29–34. <https://doi.org/10.1108/10650740410512301>
- Hargreaves, D. H. (1997). In defence of research for evidence-based teaching: A Rejoinder to Martyn Hammersley. *British Educational Research Journal*, 23(4), 405–419. <https://doi.org/10.1080/0141192970230402>
- Huber, J. A. (2004). A Closer look at SQ3R. *Reading Improvement*, 41(2), 108–112.
- Hwang, G. J., & Chang, H. F. (2011). A Formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, 56(4), 1023–1031. <https://doi.org/10.1016/j.compedu.2010.12.002>
- Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. <https://doi.org/10.1007/s13748-016-0094-0>
- Kuromiya, H., Majumdar, R., & Ogata, H. (2020). Fostering evidence-based education with learning analytics. *Educational Technology & Society*, 23(4), 14–29.
- Li, L. Y., Fan, C. Y., Huang, D. W., & Chen, G. D. (2014). The effects of the e-book system with the reading guidance and the annotation map on the reading performance of college students. *Educational Technology & Society*, 17(1), 320–331.
- Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R news*, 2(3), 18–22. <https://www.researchgate.net/publication/228451484>
- Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croatian Operational Research Review*, 7(2), 367–388. <https://doi.org/10.17535/corr.2016.0025>
- Moran, D. J., & Malott, R. W. (2004). *Evidence-based educational methods*. Elsevier Academic Press.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71. https://doi.org/10.1207/s15326985ep3801_8
- Park, H., & Kwon, H. C. (2011). Improved Gini-Index algorithm to correct feature-selection bias in text classification. *IEICE Transactions on Information and Systems*, E94.D(4), 855–865. <https://doi.org/10.1587/transinf.e94.d.855>
- Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1), 77–93. <https://doi.org/10.1023/b:amai.0000018580.96245.c6>
- Rienties, B., Cross, S., & Zdrahal, Z. (2016). Implementing a learning analytics intervention and evaluation framework: What works? *Big Data and Learning Analytics in Higher Education*, 147–166. https://doi.org/10.1007/978-3-319-06520-5_10
- Russell, R. S., & Taylor, B. W. (2008). *Operations management* (4th ed.). Prentice Hall College Div. <https://www.sciencedirect.com/book/9780750649957>

- Saks, K., Pedaste, M., & Rannastu, M. (2018). University teachers' and students' expectations on learning analytics. In M. Chang., N. S. Chen, R. Huang, K. Kinshuk, K. M. Moudgalya, S. Murthy, & D. G. Sampson (Eds.), *Proceedings of 18th International Conference on Advanced Learning Technologies* (pp. 183–187). IEEE. <https://doi.org/10.1109/ICALT.2018.00050>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Siemens, G. (2013). Learning analytics. *American Behavioral Scientist*, 57(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: A Case study in Taiwan. *International Journal of Educational Technology in Higher Education*, 17(1). <https://doi.org/10.1186/s41239-020-00186-2>
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The Current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
- West, D., Tasir, Z., Luzeckyj, A., Si Na, K., Toohey, D., Abdullah, Z., Searle, B., Farhana Jumaat, N., & Price, R. (2018). Learning analytics experience among academics in Australia and Malaysia: A Comparison. *Australasian Journal of Educational Technology*, 34(3). 122–139. <https://doi.org/10.14742/ajet.3836>
- Xing, W., & Du, D. (2018). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, 57(3), 547–570. <https://doi.org/10.1177/0735633118757015>
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2(1), 1–5. <https://doi.org/10.1016/j.caeai.2021.100008>
- Yin, C., & Hwang, G. J. (2018). Roles and strategies of learning analytics in the e-publication era. *Knowledge Management & E-Learning: An International Journal*, 10(4), 455–468. <https://doi.org/10.34105/j.kmel.2018.10.028>
- Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2018). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A Human factor approach. *International Journal of Human-Computer Interaction*, 35(4–5), 313–322. <https://doi.org/10.1080/10447318.2018.1543077>
- Zhao, F., Hwang, G. J., & Yin, C. (2021). A Result confirmation-based learning behavior analysis framework for exploring the hidden reasons behind patterns and strategies. *Educational Technology & Society*, 24(1), 138–151.

A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning

Alwyn Vwen Yen Lee^{1*}, Andrés Carlos Luco² and Seng Chee Tan¹

¹National Institute of Education, Nanyang Technological University, Singapore // ²Nanyang Technological University, Singapore // alwyn.lee@nie.edu.sg // acluco@ntu.edu.sg // sengchee.tan@nie.edu.sg

*Corresponding author

ABSTRACT: Although artificial Intelligence (AI) is prevalent and impacts facets of daily life, there is limited research on responsible and humanistic design, implementation, and evaluation of AI, especially in the field of education. Afterall, learning is inherently a social endeavor involving human interactions, rendering the need for AI designs to be approached from a humanistic perspective, or human-centered AI (HAI). This study focuses on the use of essays as a principal means for assessing learning outcomes, through students' writing in subjects that require arguments and justifications, such as ethics and moral reasoning. We considered AI with a human and student-centric design for formative assessment, using an automated essay scoring (AES) and feedback system to address issues of running an online course with large enrolment and to provide efficient feedback to students with substantial time savings for the instructor. The development of the AES system occurred over four phases as part of an iterative design cycle. A mixed-method approach was used, allowing instructors to qualitatively code subsets of data for training a machine learning model based on the Random Forest algorithm. This model was subsequently used to automatically score more essays at scale. Findings show substantial agreement on inter-rater reliability before model training was conducted with acceptable training accuracy. The AES system's performance was slightly less accurate than human raters but is improvable over multiple iterations of the iterative design cycle. This system has allowed instructors to provide formative feedback, which was not possible in previous runs of the course.

Keywords: Automated essay grading, Human-centric AI, Formative feedback, Machine learning, Ethics education

1. Introduction

Over the past decades, the deployment of Artificial Intelligence (AI) has transited from a nascent idea into an established field that is widespread and undeniably impactful on education with profound possibilities (Holmes et al., 2019). With untapped potential to create impacts by augmenting human intelligence with machine intelligence for educational research and purposes (Yang, 2021), there is also growing research on how AI can sustainably do so (Vinuesa et al., 2020). However, even though the advancement of AI entails the need to enable human welfare by improving human conditions, there remains a critical need to investigate how AI can be responsibly designed, implemented, and evaluated, especially in the field of education. Afterall, learning is inherently a social endeavor involving human interactions and not just disembodied human-machine transactions (D'Mello, 2021), rendering the need for AI designs to be approached from a humanistic perspective as human-centered AI (HAI) with consideration of human conditions and contexts (Yang et al., 2021).

This is more recently viewed to be an emergent and urgent concern, as an increasing number of functions in AI systems have already been ceded to algorithms to the detriment of human control, resulting in growing unease and loss of equitability (Sareen et al., 2020). Further, as a new-age workforce constantly evolves with a constant flux of expectations and needs, the identification of potential knowledge gaps and deficits of expertise in higher education can help support the development and implementation of AI in education (Lee, 2020). Students can remain relevant in the new reality by equipping themselves with literacies and skills to thrive in new economies while teachers adapt to new models and orientations to accommodate lifelong learning (Aoun, 2017). It is therefore unsurprising to note that a growing number of recent studies and meta-studies have utilized AI-supported systems (e.g., Garcia-Magarino et al., 2019; Lepri et al., 2021) but are focusing more on trustworthy systems with explainable layers, so that users have the opportunity to understand the reasons behind decisions. AI designs should also then consider human conditions and have a human-oriented approach when augmenting human intelligence with machine intelligence (Yang et al., 2021).

Students as future leaders will face the above-mentioned challenges as AI continues to shape society. Therefore, considering how students navigate the existent knowledge society, the study of ethical reasoning plays a key role in enhancing students' problem-solving capacity and exercising their minds in the disciplines of critical and logical thought. However, the use of AI in the domain of philosophy remains limited due to differences in

philosophical, pedagogical, and technological approaches. On one hand, it may be surprising to some that most AI work does not require any philosophy since a restricted representation has already been designed or programmed (McCarthy, 1995). On the other hand, this should not detract from the potential of using AI in studies of philosophy, of which the ease of study can allow both teachers and students in higher education to better adapt to new ways of teaching and learning. Even so, emergent societal needs such as sustainable assessment for lifelong learning will require significant shifts away from the current focus of assessment of learning (summative assessment) to assessment for learning (formative assessment) (Nguyen & Walker, 2016).

When undergoing this transition, the successful use of AI in the form of Automated Essay Scoring (AES) within the field of summative assessment of learning (Gardner et al., 2021) can offer exciting opportunities for formative assessment. Gardner and his co-authors opined that “AI in educational settings has changed little in its basic percepts and functions” and lamented that machine learning and actions have not delved far beyond intelligent analysis of large-scale data in the last decade. Thille et al. (2014) argued that large-scale assessment should benefit learners by providing continuous, multi-faceted feedback. In this regard, recent advances in AI technologies afford opportunities for formative assessment at scale, such as using machine learning to determine the quality and distribution of ideas in classroom discourse (Lee, 2021) and using trace data to dynamically give young learners immediate performance feedback in comprehension tasks (Walker et al., 2017).

To address these challenges, we attempt to answer how HAI can be designed and used for formative assessment, using processes that adjust algorithms through human contexts and social phenomena. The context of this study represents a situation that is prevalent in many foundational undergraduate courses, which are often offered to large cohorts of students. The course in this study, “Ethics and Moral Reasoning,” has an overwhelming number of student registrants, often ranging over 600 students each semester. With these students trekking through an online module that is often supervised by few instructors, several imminent problems became apparent: (1) The course has to be conducted online due to the large number of students, which further enlarges the perceived distance between the instructors and students; (2) for every assignment issued to students to gauge their understanding and progress of learning, the number of returned assignments is overwhelming for a small team of instructors to score accurately and in a timely manner; (3) providing personalized and meaningful feedback to students becomes nearly impossible; and consequently, (4) some students may not be able to grasp the importance and significance of ethics and moral education based on limited interactions with the instructors.

In this study, we use a mixed-method approach consisting of human-designed rubrics for assignment coding, peer assessment and application of machine learning as part of an automated essay scoring and feedback system for the development of ethical reasoning. In response to the challenges of courses with large enrolments that are conducted in an online format, this study attempts to answer the following research question: *To what extent can an automated essay scoring and feedback system be employed to provide formative feedback and potentially act as a surrogate for instructor interactions?*

Addressing this question will benefit the teaching and learning in courses with large enrolments, especially when more online courses are being added to the university’s offerings due to the emerging dynamics of the educational landscape and deployment of educational technologies. A caveat is that the deployment of AI, in the context of education practices and computing development, will likely not take over the role of the instructors, due to how teaching and learning happen in the classroom and the ways in which it is profoundly different from human intelligence that AI seeks to emulate (Cope et al., 2020). More importantly, tools and systems modified through this study can focus on learning from human inputs and collaborations, to support course designs that focus on improving humanistic aspects such as students’ communications and critical thinking skill development, through the provision of formative feedback that is more timely, meaningful, and actionable.

2. Background

In this section, we highlight the importance of education in ethics and moral reasoning, followed by the significance of formative feedback, an overview of several automated essay scoring and feedback systems, and lastly, our selection of method in this study.

2.1. Importance of education in ethics and moral reasoning

For an emergent knowledge society to assimilate meaningful use of AI for teaching and learning, students will need to develop ethical reasoning to enhance problem-solving capacity and exercise minds in the disciplines of

critical and logical thinking. In an ideal situation, courses in ethics put students on paths toward what Lawrence Kohlberg, a famous psychologist, termed “postconventional” moral reasoning (Rest et al., 1999). At this stage of moral reasoning, “individual judgments are now determined by self-chosen, internal principles rather than accepted from external authorities” (Vozzola, 2014, p. 29). To cultivate skills in postconventional moral reasoning, students should have ample opportunities to express their values. More importantly, they should be challenged to defend and refine their values in response to feedback from others. By participating in a university course in ethics, students are not just introduced to moral values that one or another thinker believes in, they are also challenged to reflectively cultivate their own values. They are given sufficient space and opportunities to express themselves and defend or refine their values and opinions in response to others.

In addition, as McKeachie and Svinicki (2006) noted, “values are not likely to be changed much simply by passively listening and observing a lecturer. Change is more likely in situations in which the teacher and the students reflect, listen, and learn from one another” (p. 338). In order to develop good values and live reliably by them, one needs to develop skills in moral reasoning, which is the ability to independently assess a situation, identify morally relevant considerations, and arrive at judgments about what one ought to do. Thus, in wanting to be ethical during undergraduate studies and in the society that awaits students after graduation, they have to be able to think through complex moral situations by themselves and rely on their own powers of moral reasoning. The course that students undertake in this study is an opportunity and setting that provides a sampling of such situations. For such a course, regular formative assessment and feedback provided by peers and instructors are deemed to be critical.

2.2. Significance of formative feedback for teaching and learning

Formative feedback is defined as the “information communicated to the learner that is intended to modify his or her thinking or behavior to improve learning” (Shute, 2008, p. 153) or in layman terms, is any message delivered to a learner while there is still time to adjust. Receiving feedback challenges learners’ existing beliefs and forces them to evaluate their positions. Formative feedback is not limited by fields and can be relevant in the sciences (Shavelson et al., 2008), engineering (Roselli & Brophy, 2006), the humanities and life in general (Shute, 2008). In general, formative feedback can be provided in many ways, from teachers’ written feedback to full critique sessions of an engineer’s work-in-progress (Shute, 2008).

A meta-analysis conducted by Hattie (2013) found that among all the pedagogical methods that instructors have at their disposal, the provision of formative feedback consistently yields one of the most powerful effect sizes. Formative feedback, when used for the enhancement of learning and achievement, can help instructors realign their teaching in response to learners’ needs (Jawah et al., 2004). The importance of formative feedback cannot be overstated as it motivates learners to take greater agency in their learning, and potentially provides a direction for improvement. Although feedback is among the major influences, the type of feedback and the way it is given can be differentially effective (Hattie & Timperley, 2007), such as the timing of feedback and both positive and negative impacts on learning.

When providing effective formative feedback for teaching and learning purposes, essay writing is considered one of the most useful tools for either assessing students’ learning, their ability in organizing and integrating of ideas into a knowledge artifact, or the competency of expressing oneself in writing (Valenti et al., 2003). The scoring of free-written responses such as essays, however, is a non-trivial process with inherent challenges such as the perceived subjectivity of the grading process. Hence, this problem attracted a large range of methods and techniques as solutions, including neural approaches (e.g., Taghipour & Ng, 2016), techniques such as Bayes’ theorem (e.g., Rudner & Liang, 2002), and more prevalently natural language processes involving semantic analysis (e.g., Rokade et al., 2018) to grade free form texts or essays.

From the instructors’ perspective, the availability of technology does alleviate parts of the teaching load, but there remains potential pedagogical impediment that affects instruction and assessment. For example, apart from administrative duties, instructors are still expected to handle large groups of students (i.e., lopsided student-teacher ratio) with insufficient scaffolds or tools to facilitate meaningful teaching and learning. To be responsible for the learning needs of a large group of students, it is extremely challenging for instructors to contextualize assignments, correct misconceptions, and still provide timely and accurate feedback – practices that are beneficial for students’ personal growth (Hattie & Timperley, 2007). To mitigate the severity of such issues, prior research has suggested peer reviews and evaluations as possible strategies that prompt students to complete assignments in a diligent manner (Liu & Carless, 2006). An alternative to other potential solutions, including expansion of teaching teams or leveraging on peer reviewers, is to automate the grading processes within the

course, at least partially, so that more time and space are freed up for instructors to set up well-established routines that provide feedback to students.

2.3. Automated essay scoring and feedback systems

Automated essay scoring (AES) systems attempt to accomplish part of what instructors do in assessment – evaluate students’ work and provide feedback to the students. Even as far back as several decades ago, the goal of these systems has always been to make them at least indistinguishable from human raters (Page, 1966). The eventual goal of these systems is to deliver a consistent assessment comparable with human graders. To develop an AES system, a large dataset is often split up into smaller subsets of data, with some subsets allocated for training and the remainder for validation and testing. The system firstly utilizes the marks scored by experts (which in many cases are the instructors) as labels, then attempts to generate models based on the source material (essays), before using the models to score the remaining essays in the dataset. To approach expert levels of analysis and accuracy, additional training sets labelled with expert ratings are often used in multiple passes of the training dataset, also known as epochs.

The field has developed much since Ellis Page and his colleagues developed the first AES system, Project Essay Grade (PEG), for college-level and adult writers (Page, 1966). Essentially, like many current AES systems, measures are used to approximate features of interest and describe the quality of essays designed for students and writing practice. Since then, several prominent commercial AES systems have been developed and improved, such as E-rater (Attali & Burstein, 2006), Intelligent Essay Assessor (Foltz et al., 1999), and Intellimetric (Elliot, 2003). These AES systems, similar to PEG, assist teachers in the process of essay scoring by allowing students to write and submit their work before the system provides automated feedback. The systems are mostly capable of scoring the essays and providing suggestions for improvement in targeted areas such as language, style, and sentence structure.

For example, Educational Testing Service (ETS) has used E-rater since 1999, based on intuitively small and meaningful features to score essays (Attali & Burstein, 2006). These micro-features produce a single scoring model that can be used across different assessment prompts. It also allows easier modification and upgrading of the system, so as to boost overall grading performance. Common features include grammar, word usage, sentence mechanics, style, lexical complexity, and prompt-specific vocabulary usage. Upgrades to E-rater were designed to flag anomalous and bad-faith essays, which are not scored, while scores for other essays are calculated using a weighted average of standardized feature values followed by a linear transformation to achieve the desired scale. A distinguishing factor E-rater has over other AES is the possible use of judgmental control by end-users, enabling users to determine relative weights, either by content experts or by settings weights based on similar assessments, hence preventing extreme relative weights from affecting the validity of scores.

Another example is the Intelligent Essay Assessor (IEA) (Foltz et al., 1999), which provides an alternative to using an expert training set. IEA is based on Latent Semantic Analysis (LSA) (Landauer et al., 1998). Using domain-representative texts like textbooks, articles, and samples of writing for training, LSA derives a high-dimensional semantic representation of the information within the domain by using vectors, often referring to lists or columns of scalar real numbers, to represent the words and semantic information found in the source material. Vectors may represent sources like student essays and the closer vectors are to each other, the more similar the essays are. Hence, scores for essays can then be determined by comparing each essay against all previously graded essays of similar vector weights. The result is a holistic determination of essay quality and this system can also be used as a generalized solution that extends to subjects such as psychology, biology, and history as well as ETS’s Graduate Management Achievement Test (GMAT). Past results have shown that IEA’s reliability is comparable to that of human graders, with other features including flagging anomalous essays and essays that are too similar to each other or textbooks, indicating possible instances of plagiarism. However, as much as LSA is used as a formative assessment tool, IEA is not originally designed to be a teaching tool. It compares texts based on semantic similarity, but it cannot assess creative writing or point writers toward improvement of their texts.

The Intellimetric is a proprietary intellectual asset protected by Vantage Learning (Elliot, 2003). The system parses text to flag the syntactic and grammatical structure of essays. The sentences are then tagged for parts of speech, grammatical structures, and concept expressions. Unique words and concept networks are subsequently employed for spelling recognition and grammar checking. The data is coded to form models and these models are then associated with features extracted from text and tentative scores may be assigned. Optimization is eventually performed to provide a single grade to the essay. The robustness in this system comes in the form of

using different models to grade the essays, similar to how multiple judges are employed to conduct manual essay grading.

2.4. Choice of AES system for this study

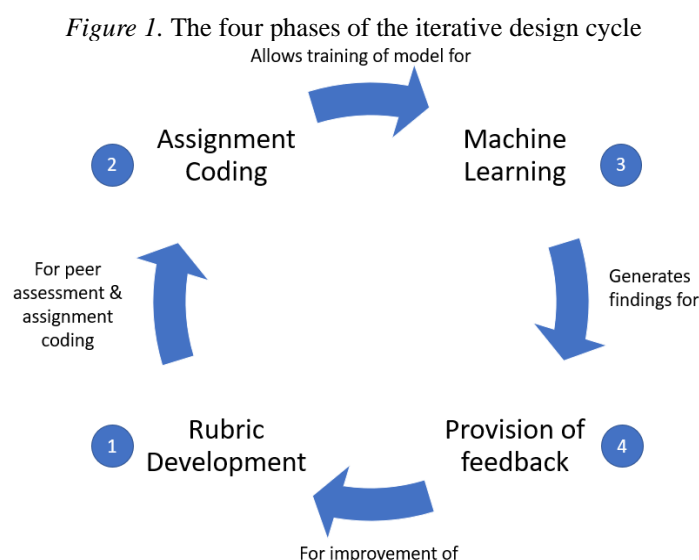
In summary, the above-mentioned are commercially-ready and established AES systems that provide many benefits to users over multiple iterations of design and improvements, but many of them are proprietary and closed source, or are platform-dependent and require conforming usage to a specific system. The goal of this study is to determine to what extent and form can an AES system exist to combine quantitative features with essay content to provide a reliable method for scoring, and potentially in place of instructors. For social scientists, since developments in algorithms are useful only to the extent that they can access the implementation (Schonlau & Zou, 2020), therefore, machine learning algorithms that are open-source and based on supervised learning models provide an intermediate solution for solving problems that are difficult to solve via conventional programs but are yet able to learn without being explicitly programmed.

This solution was sought due to the ability to monitor the performance of models and adjust parameters whenever necessary based on the accuracy of prediction. Due to the need for a score-based system, regressions are chosen to be used and among a list of regression algorithms, several studies (Ghanta, 2019; Liu et al., 2012; Schonlau & Zou, 2020) have shown that the random forest models tend to have better prediction accuracy than other regression algorithms (e.g., linear regression, logistic regression, support vector regression) over multiple sets of data. It also fits well with the iterative design cycle (Figure 1) that will be described later in the next section.

3. Method

3.1. Research design

A mixed methods approach was used, involving the analysis and evaluation of qualitative measures during peer assessment and assignment coding, and the use of quantitative methods from the machine learning application. It was part of an iterative design cycle, which is commonly a design-implement-evaluate cycle, where data and analyses in this study can inform and improve the design and scope of learning interventions during subsequent cycles. The provision of a closing feedback loop caters to how we can evaluate the broadening of the study's reach to incorporate other types of learning activities and courses. In this study, we iterated once through the cycle to illustrate the four phases that sequentially occur during the development of the AES system for an actual undergraduate course at a university. These four phases are: (1) Rubric development for peer assessment and assignment coding; (2) assignment coding by instructors; (3) machine learning application; and (4) follow-up actions in providing feedback to students. These phases are further detailed in the following subsections.



3.2. Settings and data

In this study, over 600 undergraduate students from across the university attended the course “Ethics and Moral Reasoning,” with the entirety of the course being delivered online and split into 13 sessions, also known as units in this study. These units include three major ethical theories’ utilitarianism, Kant’s deontology, virtue ethics, ethical principles for academic integrity and research, and ethics for sustainability and conservation. Students sequentially progress through the 13 units, at a pace of one unit per week throughout a semester.

The majority of the units began with a short video lecture that covered a well-defined topic in the domain of ethics and moral reasoning, followed by a short selection of readings. As part of students’ contributions to the course, each student was requested to either write a short essay (more than 100 words) to a question or to provide a short response (also in short essay format) to another student’s essay during some point in the course. Students were encouraged to contribute at least once throughout the course, which led to responses being distributed across the course units. The description and distribution of the essays in the entire dataset is shown in Table 1. These writing assignments also became an entry point for the introduction of student-centered formative feedback.

Table 1. Description and distribution of essays throughout the dataset, with no essays required for the 1st unit

Course unit ID	Number of essays	Average length of each essay (words)
2	781	204
3	193	216
4	184	216
5	52	224
6	117	223
7	357	235
8	531	184
9	99	222
10	159	217
11	108	228

In this study, course unit 2 was selected because it has the highest number of essays. The selected course unit discussed about “Utilitarianism,” which referred students to a theory of morality that prescribes actions which maximizes happiness and wellbeing of individuals. For this topic and within the specific week where data collection was conducted, students wrote a total of 781 short essays about utilitarianism or in response to their peers’ essays, using the Discussion Board page on the Blackboard learning management system (LMS).

3.3. Rubric development for peer assessment and assignment coding

The team of instructors developed a set of rubrics (see Table 2) with defined criteria to provide guidance to instructors and students during their processes of assignment coding and peer assessment respectively. The rubrics could be used by students to guide their learning from peers and aid self-reflection of own work and were also used by instructors to code and score the short essays, which in turn became the training data for developing the machine learning model. These two processes were independent and did not affect each other: the students obtained formative feedback from peers, while the instructors provided inputs for the machine learning model.

The assignment coding process was conducted by two raters to address the consistency of the proposed implementation and to also obtain a measure of interrater reliability. Due to practical reasons in needing to grade thousands of students with ten units of assignments each, the scoring system was simplified for instructors to use the rubric with the four criteria as a guideline and the theory of majority rule to provide an eventual score of 1 if the essay fulfils the majority of criteria and 0 if otherwise. Essays that received conflicting scores were returned to the pair of raters for rescoreing.

Table 2. Rubrics for peer assessment and assignment coding

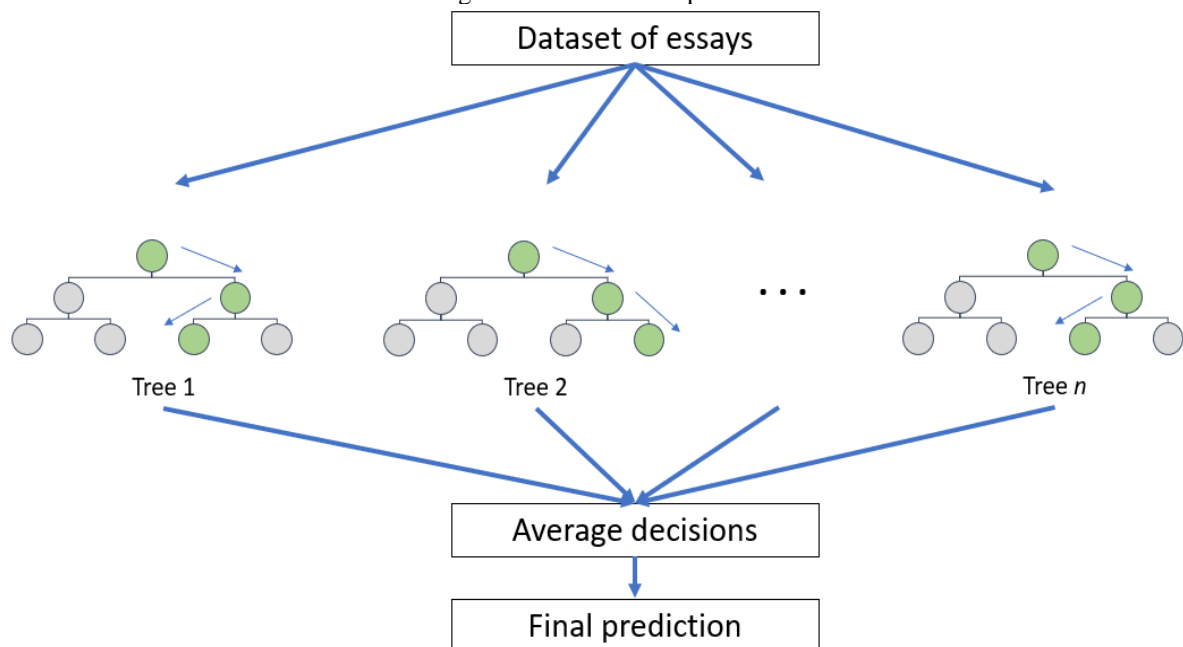
Defined criteria	Criteria grade			
	Excellent	Good	Needs improvement	Poor
Relevance – Content addresses discussion question	Very relevant to the discussion question.	Relevant to the discussion question, contains some digressing content.	Somewhat relevant to the discussion question, with some off-topic content.	Off-topic and no relevance to the discussion question.

Comprehension – Ability to accurately explain key concepts	Accurately explains key concepts necessary for responding to the question, with the use of critical keywords – e.g., utilitarianism; consequences etc.	Largely accurate in explaining key concepts with some minor flaws, with the use of the following keywords – e.g., utilitarianism; consequences etc.	Reflects major misunderstandings of key concepts, failing to refer to relevant key concepts.	Displays no understanding of key concepts.
Thesis – Central statement with at least one argument	Clear and explicit thesis statement.	Explicit thesis but not clearly stated.	Thesis is present but not clear and explicit.	Lacks a thesis statement.
Arguments and reasons for thesis	Clearly lists some pros and cons, weighed against each other to support thesis, preferably with examples.	Pros and cons are suggested but not clearly stated or are not weighed against each other to support the thesis.	Either pros or cons are provided but not both.	Lacks any effort to provide pros and cons.

3.4. Machine learning – Random Forest classifier

The instructors coded a subset of the entire dataset from the course unit “Utilitarianism,” which was then used as the training set for an open-source AES system that is modified to work with the LMS in the university. Simply put, if the AES system manages to train a model that has an acceptable level of accuracy (also known as training accuracy) based on a reasonable interrater reliability measure, the model will then be accepted for testing with the remaining parts of the entire dataset before evaluating whether the eventual model can be used for deployment in the LMS.

Figure 2. Essays are processed using multiple trees in a Random Forest algorithm before the decisions are averaged and reach a final prediction



In this study, the Random Forest classifier (Breiman, 2001) was used as a supervised learning algorithm that utilizes the ensemble learning method for regression, by building multiple decision trees and merging them in order to achieve a more accurate and stable prediction. Python was used to code the entire process and the ensemble method of seeding a forest of decision tree learners started with individual decision trees that were grown at random and acted as weak learners, with each tree presenting an outcome that was then coupled together with other trees to create a final forest. When the decisions of this forest were averaged, Random Forest

determined the weight of trees that would be used in the final model for prediction, which could then be utilized for automated scoring of unlabeled essays. This process is visually represented in Figure 2.

However, before the classifier can be implemented, several steps need to be conducted. These include firstly extracting textual data from a multitude of essays on the LMS, before running it through a spelling correcting process such as Norvig's spelling corrector (2007). This is part of preprocessing to avoid interfering with semantics and also because spelling and grammar in this study were not a major consideration in the scoring process. This was followed by feature engineering and extraction to generate multi-dimensional feature vector representations for each essay and outputting into feature arrays, before model training could take place. Random Forest was then implemented to generate a model, after which Cohen's Kappa value (Cohen, 1960) was used as a measure of agreement between the human rater and how well the model predicted using the machine learning algorithm, with correction for chance agreement.

3.5. Provision of feedback

Formative feedback was given to students in two formats. The first kind of feedback was the aforementioned score that was built into the AES system, allowing the instructors to provide a score, albeit a binary one, for every essay that was written by students. This feedback mainly serves to recognize students' effort in thinking, writing, and responding to online discussions, further encouraging them to continue sharing and discussing their ideas and thoughts with fellow peers. This was considered an improvement over the previous iterations of the course, when formative feedback was not largely available because it was impractical for a small team of instructors to consistently read and grade hundreds of short essays every week for 11 weeks throughout the semester.

The second kind of feedback was intended to be part of a larger goal in integrating meaningful feedback into the LMS. Because the essays were graded by machines throughout the semester, the instructors were able to shortlist a range of essays for deeper reading based on the scores that were generated with a level of confidence. With the ability to shortlist essays for further discussions with the students, they could gain a better grasp of the importance and significance of ethics and moral education based on their peers' work and through increased interactions with the instructors. By design, the formative feedback focused on argumentation, ethical reasoning, and critical analysis rather than looking at lower-level skills like grammar and sentence structure in the short essays. While the provision of this latter kind of feedback can be beneficial, especially when scaled with the university's learning systems, the focus in this study, however, was more on the former kind of feedback since it is critical to generate accurate results that allow the latter kind of feedback to emerge once the AES system stabilizes and performs robustly.

4. Findings and discussion

The findings mainly stem from phases two, three, and four of this study (from Figure 1). This section describes the accuracy of assignment coding, the training accuracy and validation of the model, followed by the implementation of the machine learning algorithm, and touches on the generated feedback as a result of the scoring.

4.1. Accuracy of assignment coding, training, and testing

During coding assignment of the training dataset, two raters initially coded a small subset of 20 short essays and provided reasons for the scoring of each essay, before coming together to resolve inter-rater differences. An inter-rater reliability (IRR) of 90.0% was initially obtained. The remaining differences were subsequently resolved after in-depth discussions between the two raters. After ensuring the two raters have achieved a high level of consistency with the scoring rubric, they proceeded to code another 167 short essays as part of the training set that was then used to train the model. The training accuracy from a training set of 187 short essays was 95.2%.

A 10-fold cross-validation was also conducted to evaluate the model and no overfitting was found. To test the model, 30 short essays scored by a human and the machine were compared. With the inclusion of the correction for chance agreement, the Kappa value for agreement between human raters and the machine model was 0.67, indicating substantial agreement between established labels by humans and the predictions by the algorithmic

model. These findings help to answer the research question that the model based on the Random Forest classifier in the AES system can perform similarly to a human rater, albeit with lower accuracy but with improvable performance, considering that this study consist of a single cycle of the scoring process and the algorithm's parameters can be further optimized.

4.2. Provision of formative feedback

The AES system presented scores for students' written essays during the study and although the scores were binary, they provided students with additional cues of how their work have been assessed and when used together with discussions on the online forums, students could better monitor and observe one's own activities, self-evaluate one's performance, and take actions based on the performance outcomes. These are important characteristics of self-regulated learning, important for academic performance as shown in other studies (e.g., Zimmerman & Moylan, 2009), but also form a critical part of how students engage themselves and others, with great relevance in the domain of ethics and moral reasoning.

The second form of feedback was given when the team of instructors provided comments on selected essays that they felt required a reaction or response. The reactions and responses may range from comments about well-written points or guidance on ideas and thoughts. By tapping on these examples, the general flow of ideas during the course can be better understood, similar to Lee and Tan's work (2017a; 2017b) and contribute towards a more productive discourse that benefits instructors and the students. If the essays were already deemed acceptable, no further comments were provided. Examples of the essays with respective scores and instructors' comments are shown in Table 3.

Table 3. Examples of essay excerpts with respective scores and formative feedback from instructors (if any)

Essay ID	Excerpt of essay on the topic of "utilitarianism" (word count)	Instructors' final label	Machine final score	Instructor comments
62	First and foremost, a utilitarian will have to consider the context in which why sex education is necessary before evaluating if he or she should proceed with it. Through sex education, students will learn important knowledge and insights into sex as a whole. The aim of the sex education in this case is not to reduce sexual behavior among students... but equip them with the knowledge to practice safe sex... (282 words)	1	1	Clear thesis with plausible reasons; good understanding of utilitarianism.
60	Firstly, a utilitarian would consider whether the decision of teaching sex education in a public school would be able to maximize overall well-being... For sensitivity context, certain aspects of the students such as their age and level of maturity should be taken into consideration before deciding whether sex education is suitable for them or rather what kind of sex education is more appropriate for that particular age group... (320 words)	0	0	Clear thesis but student appeals directly to claims about what is right and wrong, rather than deriving claims about right and wrong from the effects of actions on overall well-being.
121	From a utilitarian's point of view, their main aim is to maximize the well-being of the society. As such, we should take into account the possible benefits and implications of teaching sex education in public school. The main aim of sex education is to educate children on the potential issues related to sex. This is to prevent children from making wrong choices that may impact them greatly... (307 words)	1	1	[No comments from instructors]

4.3. Contributions and limitations of our study

In this study, our AES system has proved that it can automatically score and provide feedback to students of the "Ethics and Moral Reasoning" course with consideration of human factors. Although this capability has already

been demonstrated in some established systems (e.g., E-rater; Attali & Burstein, 2006) and emergent systems that use deep learning (e.g., Singla et al., 2021), these AES systems are also found to be over-stable (large essay changes cause little score variations) and over-sensitive (small essay changes cause large score variations). Our proposed HAI-influenced AES system partially negates these downsides with no overfitting.

However, a literal transfer of said system for use in other fields or algorithmic evolution into an all-encompassing type of algorithm with good accuracy is likely not possible soon. In other words, there is no one-size-fits-all algorithm that can be used without sacrificing certain aspects of accuracy, and although it is not an impossible task as demonstrated in an attempt by Olive et al. (2019), it can hardly compete with a predictive model for a specific course, such as the case in this study.

Nevertheless, it is possible to try and maintain a balance between achieving high accuracy (sensitivity and specificity) of essay scoring in a dedicated course and attempting to shift towards a slightly less accurate but general-use scoring system, particularly with human-based inputs and considerations. Although this effort will require tremendous resources to develop and maintain and likely not suit the objectives of every study or project, this limitation however will not detract from the benefits of developing AES systems for scoring large amounts of essays and HAI systems in general, allowing us to rethink and reflect on machine-based judgements.

5. Conclusions and future work

An automated essay scoring and feedback system was developed from open source to address several issues that arose from the running of an online course with large enrolments, further requiring automated assessment of students' work to better encourage meaningful teaching and learning. The study was divided into four phases and a mixed-method approach was used, with consideration of human-based inputs such as rubrics and qualitative coding of data subsets for training a machine learning model, which was then used to automatically score more essays at scale. Outputs and formative feedback in terms of essay scores and instructors' comments, which were lacking in previous iterations of the course, can then be provided to students, and possibly be used for fine-tuning the system's algorithms.

Returning to the research question, the AES and feedback system has shown to be beneficial in providing formative feedback to students, but it is still too early to decide whether this system can act as a surrogate for instructor interactions. This is because the implications and repercussions of replacing the teacher in a classroom can only be proven through multiple and sometimes longitudinal studies that provide evidence for explaining patterns of variables over time. However, it is undeniable that having an existing automated system that analyses and scores student essays does ease the load of instructors and provides instructors with more time to enhance activities in the course while gaining the ability to measure learning gains when needed, a benefit from the implementation of HAI design that considers human conditions and contexts.

As part of future work, once the AES and feedback system has been made more reliable and robust after several runs and validation processes, it can be integrated with an existing LMS to answer other interesting research questions, such as: "How much human interaction is required for students to feel their instructors are academically invested in them?" and "do students that receive automated feedback improve the quality of argumentation and decisions more than students who do not receive feedback?" These research questions will help drive vested interests to achieve "specific, measurable, agreed upon, realistic, and time-based" goals of smart AI research (Yang, 2019), that are generic enough to be understood by the public and also with wide-ranging implications that are meaningful to the masses.

Automated essay scoring, as a vital machine learning application over the last few decades, remains important to both instructors and students in providing summative and formative feedback for improving teaching and learning. The recent introduction of human-centric factors and adjustments to AES systems, however, has greatly helped to make learning visible and relevant to emergent user needs. As the need for AES becomes more imperative with the growing emphasis on remote and online learning, and with the aid of emerging techniques and technological affordances, the use of HAI designs in automated essay scoring may eventually become more widely implemented and commonplace.

Acknowledgement

This study was funded by the Startup Fund under the Centre for Research and Development in Learning (CRADLE), Nanyang Technological University, Singapore. The views expressed in this paper are the authors' and do not necessarily represent the views of the host institution. The research team would also like to thank the instructors and student participants involved in this study.

Conflict of interests

The authors confirm there are no conflicts of interests.

References

- Aoun, J. E. (2017). *Robot-proof: Higher education in the age of artificial intelligence*. MIT press.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(3). <https://ejournals.bc.edu/index.php/jtla/article/view/1650>
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Cohen, J. (1960). A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://doi.org/10.1177%2F001316446002000104>
- Cope, B., Kalantzis, M., & Sears, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, 1-17. <https://doi.org/10.1080/00131857.2020.1728732>
- D'Mello, S. (2021, July 1). *From modeling individuals to groups: It's a multimodal multiparty* [Keynote session]. Educational Data Mining 2021, Paris, France.
- Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A Cross-disciplinary Perspective* (pp. 71-86). Routledge.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2). <http://imej.wfu.edu/articles/1999/2/04/index.asp>
- Garcia-Magarino, I., Muttukrishnan, R., & Lloret, J. (2019). Human-centric AI for trustworthy IoT systems with explainable multilayer perceptrons. *IEEE Access*, 7, 125562-125574.
- Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?". *Journal of Computer Assisted Learning*, 37(5), 1207-1216. <https://doi.org/10.1111/jcal.12577>
- Ghanta, H. (2019). *Automated essay evaluation using natural language processing and machine learning* (Unpublished master's thesis). Columbus State University, Columbus, GA. https://csuepress.columbusstate.edu/theses_dissertations/327/
- Hattie, J. (2013). *Visible learning: A Synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J., & Timperley, H. (2007). The Power of feedback. *Review of educational research*, 77(1), 81-112. <https://doi.org/10.3102%2F003465430298487>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education*. Center for Curriculum Redesign.
- Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). Enhancing student learning through effective formative feedback. *The Higher Education Academy*, 140, 1-40.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284. <https://doi.org/10.1080/01638539809545028>
- Lee, A. V. Y. (2020). Artificial intelligence in education (AIED). In H.-J. So, M. M. Rodrigo, J. Mason, & A. Mitrovic (Eds.), *Proceedings of the 28th International Conference on Computers in Education (ICCE), Volume 2* (pp. 749-750). Asia-Pacific Society for Computers in Education.

- Lee, A. V. Y. (2021). Determining quality and distribution of ideas in online classroom talk using learning analytics and machine learning. *Educational Technology & Society*, 24(1), 236-249.
- Lee, A. V. Y., & Tan, S. C. (2017a). Discovering dynamics of an idea pipeline: Understanding idea development within a knowledge building discourse. In W. Chen, J.-C. Yang, A. F. Mohd Ayub, S. L. Wong, & A. Mitrovic (Eds.), *Proceedings of the 25th International Conference on Computers in Education (ICCE) 2017* (pp. 119-128). Asia-Pacific Society for Computers in Education.
- Lee, A. V. Y., & Tan, S. C. (2017b). Understanding idea flow: Applying learning analytics in discourse. *Learning: Research and Practice*, 3(1), 12-29. <http://dx.doi.org/10.1080/23735082.2017.1283437>
- Lepri, B., Oliver, N., & Pentland, A. (2021). Ethical machines: The Human-centric use of artificial intelligence. *IScience*, 24(3), 102249. <https://doi.org/10.1016/j.isci.2021.102249>
- Liu, N. F., & Carless, D. (2006). Peer feedback: The Learning element of peer assessment. *Teaching in Higher education*, 11(3), 279-290. <https://doi.org/10.1080/13562510600680582>
- Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer.
- McCarthy, J. (1995). What has AI in Common with philosophy? In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2041-2044).
- McKeachie, W. J., & Svinicki, M. (2006). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers*. Houghton Mifflin Company.
- Nguyen, T. T., & Walker, M. (2016). Sustainable assessment for lifelong learning. *Assessment & Evaluation in Higher Education*, 41(1), 97-111. <https://doi.org/10.1080/02602938.2014.985632>
- Norvig, P. (2007). *How to write a spelling corrector*. <http://norvig.com/spell-correct.html>
- Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A Quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies*, 12(2), 171-183. <https://doi.org/10.1109/TLT.2019.2911068>
- Page, E. B. (1966). The Imminence of grading essays by computer. *Phi Delta Kappan*, 48, 238-243.
- Rest, J. R., Thoma, S. J., & Bebeau, M. J. (1999). *Postconventional moral thinking: A Neo-Kohlbergian approach*. Psychology Press.
- Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018). Automated grading system using natural language processing. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1123-1127). IEEE. <https://doi.org/10.1109/ICICCT.2018.8473170>
- Roselli, R. J., & Brophy, S. P. (2006). Experiences with formative assessment in engineering classrooms. *Journal of Engineering Education*, 95(4), 325-333. <https://doi.org/10.1002/j.2168-9830.2006.tb00907.x>
- Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2). <https://ejournals.bc.edu/index.php/jtla/article/view/1668>
- Sareen, S., Saltelli, A., & Rommetveit, K. (2020). Ethics of quantification: Illumination, obfuscation and performative legitimization. *Palgrave Communications*, 6, 20. <https://doi.org/10.1057/s41599-020-0396-5>
- Schonlau, M., & Zou, R. Y. (2020). The Random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3-29. <https://doi.org/10.1177%2F1536867X20909688>
- Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A Collaboration between curriculum and assessment developers. *Applied Measurement in Education*, 21(4), 295-314. <https://doi.org/10.1080/08957340802347647>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189. <https://doi.org/10.3102%2F0034654307313795>
- Singla, Y. K., Parekh, S., Singh, S., Li, J. J., Shah, R. R., & Chen, C. (2021). AES systems are both overstable and oversensitive: Explaining why and proposing defenses. *PsyArXiv*. <https://doi.org/10.48550/arXiv.2109.11728>

- Taghipour, K., & Ng, H. T. (2016). A Neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891). Association for Computational Linguistics.
- Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The Future of data-enriched assessment. *Research & Practice in Assessment*, 9, 5-16.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2(1), 319-330. <https://www.learntechlib.org/p/111481/>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmar, M., & Nerini, F. F. (2020). The Role of artificial intelligence in achieving the sustainable development goals. *Nature Communication*, 11(233), 1-10. <https://doi.org/10.1038/s41467-019-14108-y>
- Vozzola, E. (2014). *Moral development: Theory and applications*. Routledge.
- Walker, E., Wong, A., Fialko, S., Restrepo, M. A., & Glenberg, A. M. (2017). EMBRACE: Applying cognitive tutor principles to reading comprehension. In *International Conference on Artificial Intelligence in Education* (pp. 578-581). Springer, Cham.
- Yang, S. J. H. (2019, December 2-6). Precision education: New challenges for AI in education [Conference keynote]. In *Proceedings of the 27th International Conference on Computers in Education (ICCE)* (pp. XXVII-XXVIII). Asia-Pacific Society for Computers in Education.
- Yang, S. J. H. (2021). Guest Editorial: Precision education – A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of Metacognition in Education* (pp. 311-328). Routledge.

Feasibility and Accessibility of Human-centered AI-based Simulation System for Improving the Occupational Safety of Clinical Workplace

Pin-Hsuan Wang^{1,2}, Anna YuQing Huang³, Yen-Hsun Huang⁴, Ying-Ying Yang^{1,4*}, Jiing-Feng Lirng⁴, Tzu-Hao Li^{4,5}, Ming-Chih Hou^{2,4}, Chen-Huan Chen^{1,4}, Albert ChihChieh Yang^{2,4}, Chi-Hung Lin⁴ and Wayne Huey-Herng Sheu^{2,4}

¹Department of Medical Education, Clinical Innovation Center, Medical Innovation and Research Office, Taipei Veterans General Hospital, Taipei, Taiwan // ²Taipei Veterans General Hospital, Taipei, Taiwan // ³Computer Science & Information Engineering, National Central University, Taoyuan City, Taiwan // ⁴College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan // ⁵Division of Allergy, Immunology, and Rheumatology, Department of Internal Medicine, Shin Kong Wu Ho-Su Memorial Hospital // karenwang0607@gmail.com // anna.yuqing@gmail.com // michaelyhhuang@gmail.com // yanggy@vghtpe.gov.tw // jflirng@vghtpe.gov.tw // pearharry@yahoo.com.tw // mchou@vghtpe.gov.tw // chenc101@gmail.com // accyang@nycu.edu.tw // linch.ym@gmail.com // whhsheu@vghtpe.gov.tw

*Corresponding author

ABSTRACT: Medical personnel need to learn occupational safety knowledge in clinical workplaces, not only to ensure their own safety, but also to further ensure patients safety. Based on Human-centered artificial intelligence (HAI) technology, this study will provide HAI-based occupational safety training system for two training topics, Needle Stick/Sharps Injury (NSSI) prevention and appropriate Clinical Waste Management (CWM). From April 2018 to December 2021, this clinical occupational safety HAI training is used by 342 medical personnel (doctors and non-doctors). This study aims to investigate the learning performance and effectiveness including decreasing anxiety and increasing mastering level of users. This study shows that, for the first-time and feel-friendly users of this HAI training system, not only can they achieve significant learning improvement, but they can also effectively decrease their anxiety and increase their mastery level of clinical work safety knowledge and skill. In terms of learning performance and effectiveness, this study found that doctors are significantly benefited by the HAI training system in contrast to non-doctors.

Keywords: Clinical waste management, Needle stick sharp injury, Virtual reality

1. Introduction

Effective and safe patient care can only be provided if the safety of medical personnel, which includes doctor and non-doctor, are maintained. Because new medical personnel are susceptible to inappropriate Clinical Waste Management (CWM) and Needle Stick and Sharp Injuries (NSSIs). NSSI and CWM workplace safety issues are frequently discussed in hospital. Proper CWM and NSSI prevention are essential trainings for ensuring occupational safety in the clinical workplace (Gao et al., 2017; Markovic-Denic et al., 2011). In general, most of the NSSI and CWM trainings are using handouts, lectures, slides, face-to-face discussions posters, mannequin-based simulation (Ozder et al., 2013; Merandi & Williams, 2017).

Due to time and labor costs, trainings through handouts, lectures, slides, discussions, or posters are generally expensive to repeat. However, the CWM and NSSI trainings need to be available without time and space limitation so that all medical personnel can have the opportunity to repeat trainings. Therefore, hands-on learning of CWM and NSSI in HAI simulation environment may provide a solution to this challenge. Different from previous learning systems, users could gain knowledge and learn skills via human-centered learning methods by creating a personalized and self-paced tutorial (Yang, 2021), leading to a more effective medical education, especially in the occupational safety of the clinical environment.

Self-directed learning is defined as the process by which individuals guide their own learning and thereby become lifelong learners (Patterson et al., 2002; Robinson & Persky, 2020). Brockett and Hiemstra (2018) proposed that the process of self-directed learning includes four steps: planning, setting goals, selecting learning resources and re-examining the learning process, which are all completed by learners in their own learning speed. For medical personnel, self-directed learning can make them take more responsibility for the choice of learning strategies or the monitoring of self-efficacy, and then making a positive impact on their future work attitudes. Since virtual reality technology has the characteristics of instant feedback and immersive virtual environment, it is regarded as the perfect field for self-directed learning (Rozinaj et al., 2018). For this reason,

this study aims to construct a HAI-based occupational safety training system including NSSI prevention and appropriate CWM via VR technology.

A well-designed learning platform helps to build a medical education training course. Meanwhile, effectiveness including decreased anxiety and increased mastering level are important parameters to judge the quality of such learning platform. During the learning process, anxiety not only has a great impact on the learning performance of learners, but also further impairs their concentration and memory (Gibelli et al., 2019; Yang et al., 2018). For medical personnel, a high level of concentration is required to provide safe and effective patient care. Thus, this study uses two parameters, decreasing anxiety and increasing mastering level, to measure the effectiveness of HAI-based occupational safety training systems and explore following research questions:

RQ1: Exploring the impact of HAI -based training programs on the learning performance, mastery level, and user anxiety of different HAI experience groups in clinical workplace safety knowledge.

RQ2: Exploring the impact of HAI -based training programs on the learning performance, mastery level, and user anxiety of different medical personnel (doctor and non-doctor) in clinical workplace safety knowledge.

2. Related works

2.1. The occupational safety issues of clinical workplace in hospital

Needle stick and sharp injuries (NSSIs) are illustrated as percutaneous piercing wound, caused accidentally by medical or laboratory devices and appliances, such as needles, ampules, injectors, lancets, broken glass fragments, scalpels, shredded intravenous cannulation devices. Medical personnel are at high risk of needle stick or sharps injuries due to the need for repeated patient contact and various nursing behaviors in the clinical workspace of a hospital. Norsayani and Hassim (2003) further suggested that the risk of NSSI in the workplace of new medical personnel are 3 times than others. In addition, 27-40% of new medical personnel had experience NSSI during training (Wicker et al., 2008; Ghasemzadeh et al., 2015; Sharma et al., 2010). These research highlighted the necessity of NSSI education training program to improve workplace safety of new medical personnel in the hospital.

The COVID-19 pandemic has led to a surge in demand for personnel protective equipment such as gloves and masks, leading to a discussion of global waste management (Kalantary et al., 2021). With the absence of separate containers for management masks and gloves, the risk of infection from clinical waste may be substantially increased. To mitigate the risk of aerosol spread, medical personnel must consciously sort medical wastes to the correct category, which is why Kalantary et al. (2021) proposed the need for proper management of large amounts of waste in healthcare workspace. Which is to say, being able to classify the wastes and dispose them into the right collection site is a very important skill for medical personnel (Letho et al., 2021). However, a systematic review illustrated that medical personnel's confidence and familiarity toward managing clinical waster appropriately were not as expected (Ananth et al., 2010; Abebe et al., 2017; Joshi et al., 2015; Peng et al., 2020; Yazie et al., 2019).

Both general waste and hazardous waste should be properly classified according to the source of their generation (Akkajit et al., 2020). According to previous studies, hazardous clinical waste in inappropriate clinical waste management (CWM) outnumbers the proportion of general waste (Chartier et al., 2014; Hayleeyesus & Cherinete, 2016). The reason for this may be that general waste may be polluted by the hazardous waste carelessly, creating more hazard waste. It's evident that the wrong management process will increase the amount of hazardous waste. For example, contaminated needles and syringes have the potential to cause greater pollution throughout hazardous recycling and repackaging (Askarian & Malekmakan, 2006; Maina, 2018). Therefore, clinical waste management (CWM) training courses for new medical personnel are important.

Hospitals are the main field for disease treatment and health care, so the safety and hygienic requirements are relatively higher. Unfortunately, hospitals are also high-risk workplaces for occupational injury among various occupational fields. Therefore, the two most important workplace safety issues for new doctors in hospitals are medical waste management (CWM) and needle stick and sharps injury (NSSI) prevention education. In an attempt to ensure and maintain the safety of medical personnel, occupational safety training programs in hospital clinical workplaces are imperative.

2.2. The human-centered artificial intelligence is a liable design in CWM and (NSSI) prevention in our study

Artificial intelligence (AI) is a scientific principle that concentrates on creating and presenting computer algorithms that are usually designed for speeding up procedure and reducing mistakes (Hassani et al., 2020). Studies showed that, compared to conventional learning methods, AI-based learning is more likely to increased effectiveness of learning mastering level and decreased anxiety to certain fields for medical students or residents (Paranjape et al., 2019). However, in consideration of liability, trustworthy and explainability, a human-centered artificial intelligence (HAI) is now highly recommended in this field. This also applies to the training at the clinical workplace among medical personnel according to previous studies (Shneiderman, 2020).

Besides high flexibility and convenience, the virtual reality learning platform can provide 24/7 online training courses, freeing from the limitation of location, time, and personnel, achieving the purpose of training anytime, anywhere. Khunger and Kathuria (2016) also confirmed that the simulation system could help medical students learn suturing skills effectively. For medical personnel, proper clinical waste management (CWM) and prevention of NSSI are essential skills to ensure occupational safety in the clinical workplace. For these reasons, this study implemented a HAI-based occupational safety training program for CWM and NSSI units to help hospital personnel learn safety knowledge of clinical workplace.

2.3. Self-directed learning in medical education

Knowles (1975) defined self-directed learning as the process by which individuals learn independently without the help of others. In this learning process, learners will select learning resources and implement learning strategies according to their own learning needs, so as to learn knowledge independently, and finally evaluate their own learning performance. Brockett and Hiemstra (2018) proposed that the process of self-directed learning includes four steps: planning, setting goals, selecting learning resources and re-examining the learning process, which are all completed by learners in their own learning speed. Since learning goals are set by learners themselves in the process of self-directed learning, learners should be able to formulate clear, specific, and well-structured learning goals to reduce the challenges students face when learning knowledge independently. The learner's self-motivation and the learning strategies adopted will directly affect the success of self-directed learning. Therefore, the degree of self-directed learning will depend on the learner's attitude and ability.

With the rapid development of biomedical knowledge and medical technology systems, the knowledge acquired by medical schools can no longer meet the needs of hospitals in the medical field, which is also the main reason for the needs of Continuous Medical Education (CME) (Simpkin & Walesby, 2017). However, the learner's self-directed learning ability has a great influence on the learning outcomes of CME. For medical staff in hospitals, it is not only necessary to become lifelong learners, but also to develop self-directed learning (SDL) as a core skill (Ricotta et al., 2021). It can be seen that the importance of SDL in medical education.

In view of the importance of lifelong learning and self-directed learning skills for medical personnel to learn medical knowledge, more and more medical professional associations, such as the Medical Council of India (MCI), the World Federation of Medical Education (WFME) emphasis on developing the learning ability of medical staff through self-directed learning (Buch et al., 2021; Ricotta et al., 2021). Based on the theoretical concept of self-directed learning, Ricotta et al. (2021) were further proposed the framework of self-directed learning in medical education (SDL-ME). Self-directed learning in medical education (SDL-ME) focuses on the conceptualization of core attributes of medical professional identity. In the context of mutual social responsibility of medical staff and patients, medical knowledge skills and attitudes need to grow over time to implement appropriate medical care.

3. Methodology

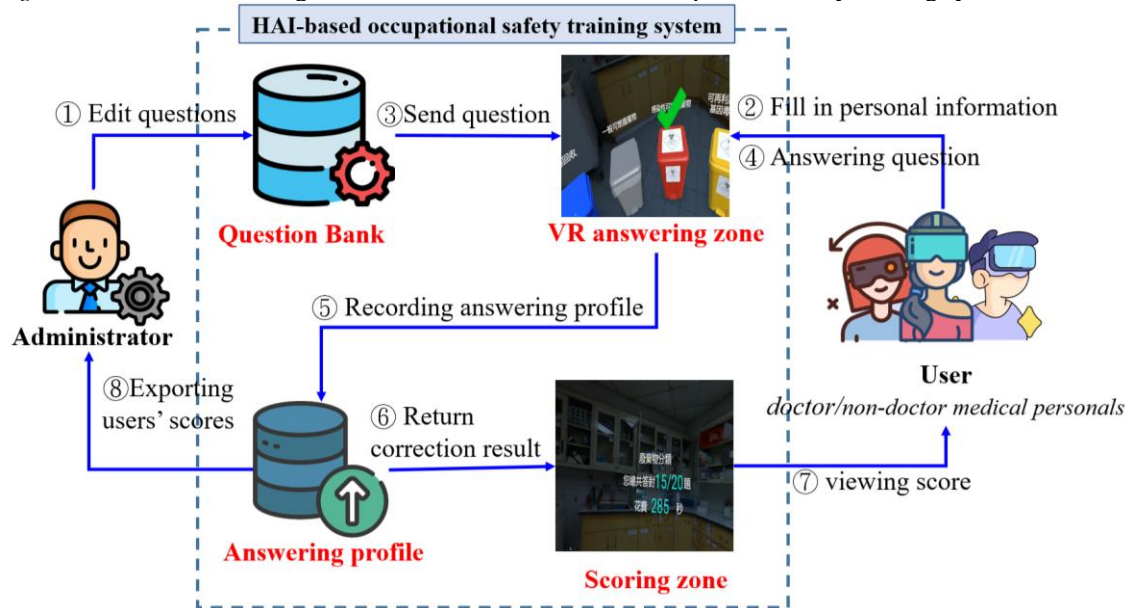
3.1. Participates

We conducted this prospective study in a 2800-bed 6000-staff medical center and teaching hospital in Taipei, Taiwan from April 2018 to December 2021. A total of 342 users, including doctors and non-doctors, were recruited for this study and randomly assigned to either NSSI prevention or CWM units of a HAI-based occupational safety training program. The Needle Stick and Sharps Injury prevention (NSSI) and Clinical Waste Management (CWM) units recruited 251 and 91 users, respectively.

3.2. HAI-based occupational safety training system

In hospitals, basic occupational training topics for clinical workplace safety mainly include needle stick/sharp injury (NSSIs) prevention and appropriate clinical waste management (CWM), usually in the form of lecture guides or demonstrations. However, due to the restriction of time and venue as well as the ever-changing schedule of trainees, it is difficult to train all medical personnel simultaneously. As a result, this study developed a HAI-based occupational safety training system, which provides a 24-hour training approach for the whole hospital and optimizes training effect based-on the real-time evaluation and feedbacks provided by the system. Figure 1 is the schematic diagram for the flow of HAI-based occupational safety training system. To elaborate, HAI-based occupational safety training system distinguished from other similar training systems by creating a precision approach, rather than a one-size-fits-all model. The precision approach provided users with personal and self-paced tutorials (Yang, 2021). It could concentrate and target on the users' weak points, leading to a safer clinical occupational environment.

Figure 1. The schematic diagram for the flow of HAI-based occupational safety training system

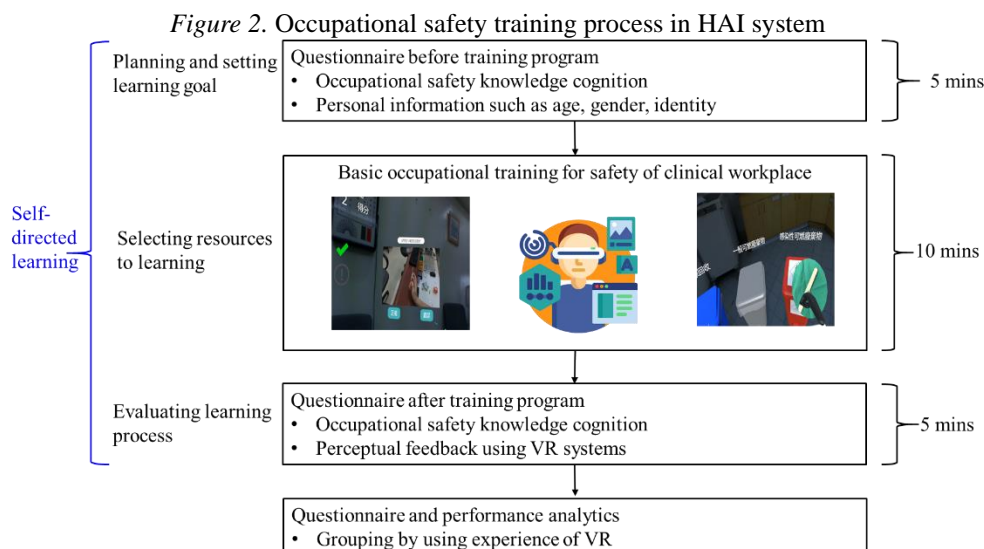


Overall, there are four components in the flow of the HAI-based occupational safety training system including question bank, VR answering zone, scoring zone, and answering profile. The user will go through the following 8 steps to partake in the training program in the HAI-based occupational safety training system. The administrator first edits 20 training questions for each of the NSSI and CWM modules in step 1. The user fills in personal information in step 2 to log in the system. Then, the question bank can start sending questions to users in step 3. The user can start the training program and answer questions in the answering zone in step 4. In order to provide training for all medical personnel, the users of this system include doctors and non-doctors. During the training process, the system transmits and stores the user's answers in the answering profile database step 5. After the user completes the training program, the task of step 6 is to extract the correction results from the answering profile database and present it in the scoring zone. Finally, in step 7, users can view the correction results on the score zone, including the number of correct answers and the time it took to answer. The main task of step 8 is to export scores of all users, so that the hospital administrator can inspect the training effect of the users.

3.3. Experiment design

Due to persistent, immersive, and highly interactive features of 3D virtual reality, it has gradually become a popular new online learning environment among educational courses (Lin & Lan, 2015). Therefore, we tried to integrate 3D virtuality with human-centered artificial intelligence. In order to achieve labour cost reduction of clinical safety training courses in hospital institutions, this study constructed a HAI-based occupational safety training system. Figure 2 is the HAI system occupational safety training process diagram. It is worth noticing that users need to fill out a questionnaire on occupational safety knowledge and personal information such as age, gender, and identity before HAI training in the HAI-based occupational safety training system for NSSI and CWM. The user identities in this study include doctor and non-doctor. The questionnaire first collects users'

basic personal information, then explains the planning and learning goals in the HAI system occupational safety training process. Next, users need to spend about 10 minutes in the HAI system to learn the clinical workplace safety knowledge about NSSI and CWM. In order to get user's self-evaluating feedback after completing the training program, users will also be asked to fill in a feedback form including occupational safety knowledge cognition and HAI system using experience. In the end, this study will classify users by HAI using experience then analyse questionnaire results and training performance.



The proposed HAI-based occupational safety training program includes CWM and NSSI prevention learning topics. Hazard waste includes toxic chemicals, pharmaceuticals, medical devices waste, radioactive substances, body fluids, discarded sharps, non-sharp, and blood. Since waste can be contaminated, infectious or dangerous, it is important to place waste in the correct storage location. Therefore, for the CWM topic, users were asked to recognize 12 random clinical wastes. The HAI-based occupational safety training program is designed to allow the trainee pick up virtual items and place them in the correct bin. Needle stick and sharp injuries (NSSIs) are illustrated as percutaneous piercing wound, caused accidentally by medical or laboratory devices and appliances, such as needles, ampules, injectors, lancets, broken glass fragments, scalpels, shredded intravenous cannulation devices. NSSI is the most common cause of workplace injuries for healthcare professionals worldwide. For the NSSIs prevention topic, users will face 12 random scenarios, which including safe/unsafe behaviours with and without universal precaution for NSSIs prevention.

3.4. Instruments

In the HAI-based occupational safety training system, users' occupational safety knowledge cognition and learning performance will be measured through system- and self-assessments, respectively. According to occupational safety knowledge measured by the system, the proposed HAI-based occupational safety training system will calculate the number of correct answers and answering time according to the user's answering profile, so as to evaluate the user's learning performance of occupational safety knowledge. As for self-assessment occupational safety knowledge, this study will ask users to self-assess their occupational safety knowledge before and after training using a four-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree).

This study will also ask users to self-assess their thoughts about the effectiveness of the HAI-based occupational safety training system through a four-point Likert scale. This study will also explore whether the system can decrease users' anxiety about occupational safety, and whether it can help users master the CWM and NSSIs skills. Table 1 listed the detailed description of the questionnaire used in this study for system- and self-assessments of knowledge, decreasing anxiety, increasing mastering level, HAI friendly, HAI experience. To explore the differences of the obtained training performance among different groups, this study will group users according to their HAI using experience. HAI using experience includes whether it is their first-time using HAI system and whether the HAI interface is user-friendly.

Table 1. Description of the questionnaire items used in this study (knowledge of system assessment, knowledge of self-assessment, anxiety, help on mastering level, HAI friendly, HAI experience)

Scale	Variable	Description of item	Response format
Learning performance (Occupational safety knowledge)	Systematic-assessment	Number of correct answers and time to answer calculated from the answering profile after completing the training program.	Values calculated by the HAI system.
	Self-assessment	Occupational safety knowledge level for CWM and NSSI units.	4: very understanding 3: understanding 2: not understanding 1: not understanding at all.
Effectiveness	Decreasing anxiety	This HAI-based occupational safety training program can effectively decrease the anxiety of occupational safety.	4: strongly agree 3: agree 2: disagree 1: strongly disagree.
	Increasing mastering level	This HAI-based occupational safety training program can effectively help you to increase the mastery level in the CWM and NSSIs skills.	
HAI system experience	HAI friendly	For the presentation of occupational safety knowledge, HAI systems are better than presentations of papers or lectures.	1: the user considers that HAI system is better than the lecture or paper presentation. 0: the user considers that HAI system is not better than the lecture or paper presentation.
	HAI experience	Are you using the HAI system for the first time?	1: the user is using the HAI system for the first time. 0: the user has used the HAI system before.

4. Experimental results and discussions

4.1. Exploring the impact of HAI-based training programs on different groups in clinical workplace safety knowledge

To explore the impact of HAI on CWM and NSSIs skills for different groups, this study divided users into groups according to two factors: whether they had the experience of using HAI or whether they felt that the HAI interface was user-friendly. The $G_Y(\text{Experience})$ and $G_N(\text{Experience})$ groups represented users with or without HAI experience, respectively. The $G_Y(\text{Friendly})$ and $G_N(\text{Friendly})$ groups represented users feeling the HAI interface user-friendly or not user-friendly. The total number of users in this experiment was 342, with 23 and 319 users in the $G_Y(\text{Experience})$ and $G_N(\text{Experience})$ groups, respectively, and 313 and 29 users in the $G_Y(\text{Friendly})$ and $G_N(\text{Friendly})$ groups, respectively. The proportion of first-time users in $G_N(\text{Experience})$ group was as high as 91.5%, and the proportion of $G_Y(\text{Friendly})$ users was as high as 91.5%. It means that most of the users had no previous experience with HAI-based training programs. The data indicated that though the majority of users felt the HAI interface user-friendly, the prevalence of usage of HAI technology in the hospital field courses remained low. However, considering the high occupational risk in the hospital field, this study aimed to create a HAI-based occupational safety training program to help users learn clinical workplace safety knowledge about NSSI and CWM.

This study used self-assessment and HAI system-based assessment to evaluate the learning performance of CWM and NSSI skills. For the self-assessment, this study conducted a four-point Likert scale pre-test and post-test on occupational safety knowledge before and after the HAI training program. For the HAI system-based assessment, this study calculated the number of correct answers in the training program as another learning performance indicator for clinical safety knowledge. Table 2 showed the independent sample *t*-test results of user self-assessment and HAI system assessment.

According to the independent sample *t*-test results of the self-assessment learning performance of the two different HAI experience groups shown in Table 2, the occupational safety knowledge pre-test of users in the $G_Y(\text{Experience})$ group was significantly higher than that of the users in the $G_N(\text{Experience})$ group ($t = 2.82, p$

< .01), but G_Y(Experience) and G_N(Experience) groups had no significant difference in the post-test of occupational safety knowledge ($t = -1.01, p > .05$). Although the pre-test of users in the G_N(Experience) group were lower than the users in the G_Y(Experience) group before training, the post-test learning performance in the G_N(Experience) group and the G_Y(Experience) group were similar. It meant that users in the G_N(Experience) group benefited more from the HAI-based training program. For the independent sample t -test results of the HAI system-based assessment of learning performance shown in Table 2, there were no significant difference between the G_Y(Experienced) and G_N(Experienced) groups in the number of correct answers ($t = .12, p > .05$). This meant that the average numbers of correct answers did not differ between the G_N(Experience) and G_Y(Experience) groups. From the descriptions above, the HAI-based assessment evaluation results showed that the users in the G_Y (Experience) group had similar learning performance with the G_N (Experience) group, but the self-assessment evaluation results showed that the users in the G_N (Experience) group learned more from the HAI-based training program.

Table 2. The independent sample t -test results of learning performance for occupational safety knowledge

Variables	Groups	#	Self-assessment				HAI system-based assessment: Correct answers	
			Pre-test		Post-test		Mean/Std.	t value
HAI Experience	G _Y (Experience)	23	3.30/.65	2.82**	3.57/.79	-1.01	10.57/4.48	.12
	G _N (Experience)	319	2.80/.85		3.73/.54		10.66/3.74	
HAI Friendly	G _Y (Friendly)	313	2.86/.83	-1.87	3.77/.49	-3.06**	10.80/3.80	2.32*
	G _N (Friendly)	29	2.55/.99		3.24/.91		9.10/3.27	

Note. # indicates the number of users in the group. * $p < .05$; ** $p < .01$.

For the self-assessment learning performance results of the two different HAI friendly groups shown in Table 2, the G_Y(Friendly) group and the G_N(Friendly) group had no significant difference in the pre-test values of occupational safety knowledge ($t = -1.87, p > .05$), but there were significant differences in the post-test values ($t = -3.06, p < .01$). The G_Y(Friendly) group and the G_N(Friendly) group had the same occupational safety knowledge before the training program, but the occupational safety knowledge of the G_Y(Friendly) group was significantly higher than that of the G_N(Friendly) group after the training program. For the results of the HAI system-based assessment learning performance shown in Table 2, there was a significant difference ($t = 2.32, p < .05$) in the number of correct answers between the G_Y(friendly) and G_N(friendly) groups. This meant that users in the G_Y(friendly) group had more correct answers than users in the G_N(friendly) group within similar time(seconds). In conclusion, users in the G_Y(friendly) group had better self-assessment learning performance, while the G_Y(friendly) group also answered more correct answers within the same time in the HAI-system based assessment.

Table 3. The independent sample t -test results of the decreasing anxiety and increasing mastery level in the CWM and NSSIs skills

Variable	#	Decreasing anxiety		Increasing mastery level	
		Mean/Std.	t value	Mean/Std.	t value
HAI Experience	G _Y (Experience)	23	2.61/.58	.27	1.07
	G _N (Experience)	319	2.65/.63		
HAI friendly	G _Y (Friendly)	313	2.72/.54	5.71***	4.82***
	G _N (Friendly)	29	1.79/.86		

Note. # indicates the number of users in the group. *** $p < .001$.

Based on users' perception of clinical workplace safety knowledge, Table showed the independent sample t -test results of decreasing anxiety and increasing mastering level. After the HAI-based training program, the G_Y(Experience) and G_N (Experience) groups had no significant difference in anxiety reduction ($t = .27, p > .05$) and mastering level ($t = 1.07, p > .05$). Nevertheless, there were significant differences between the G_Y(Friendly) and G_N(Friendly) groups in decreasing anxiety ($t = 5.71, p < .001$) and mastering level ($t = 4.82, p < .001$). This result indicated that for clinical workplace safety knowledge, decreasing anxiety and increasing mastery level of clinical workplace safety were not affected by whether users had HAI experience. In addition, users in the G_Y (Friendly) group would be able to learn better through HAI-based training programs, which not only effectively decreased the anxiety of clinical workplace safety, but also increased their mastering level of clinical safety knowledge. Based on the HAI experience and user-friendly groupings, abbreviated results were shown in Tables 2-3 and replied to RQ1 (exploring the impact of HAI-based training programs on the learning performance, mastery level, and user anxiety of different HAI experience groups in clinical workplace safety knowledge). The HAI-based training programs could effectively improve the learning performance of first-time, friendly user

groups on clinical workplace safety knowledge. In addition, for users who felt friendly, the HAI-based training program could also effectively decrease the anxiety of clinical work safety and increase the user's mastery of clinical work safety.

4.2. Exploring the impact of HAI-based training programs on different medical populations in clinical workplace safety knowledge

Based on discussions in RQ1, first-time or feel-friendly users would acquire more knowledge when learning in the proposed HAI-based clinical safety training program. Since the medical personnel in the hospital included doctors and non-doctors, this study explored the differences in the learning performance of different medical personnel between first-time and feel-friendly users. In Table 4, most of the first-time (237/319, 74.3%) or feel-friendly (231/313, 73.8%) user groups are doctors. For first-time users who belonged to the G_N (Experience) group, there was a significant difference in the number of correct answers between doctors and non-doctors ($t = 2.38, p < .05$). Doctors correctly answered 10.92 questions in average, which was significantly higher than 9.90 questions answered by non-doctors. There was also a significant difference in the number of correct answers between doctors and non-doctors for users who belonged to the G_Y (friendly) ($t = 2.65, p < .01$). Doctors answered 11.10 questions correctly in average, which was also significantly higher than 9.95 questions answered by non-doctors.

For the first-time users in the G_N (Experience) group, there were no significant differences between doctors and non-doctors in decreasing anxiety ($t = -.82, p > .05$) and increasing mastery levels ($t = -.18, p > .05$). Besides, for the feel-friendly users in the G_Y (Friendly) group, there were also no significant differences between doctors and non-doctors in decreasing anxiety ($t = -1.79, p > .05$) and increasing mastery levels ($t = -.55, p > .05$). Results above implied that both doctors and non-doctors in the first-time group and in the feel-friendly group in the HAI-based clinical safety training program felt they've been receiving equal assistance with decreasing anxiety and increasing mastery level.

Table 4. The independent sample *t*-test results of the user perception and HAI system-based assessment

Group	Identity	#	User perception				HAI system-based assessment:	
			Decreasing anxiety		Increasing mastery level		Correct answers	
			Mean/Std.	<i>t</i> value	Mean/Std.	<i>t</i> value	Mean/Std.	<i>t</i> value
G_N (Experience)	Doctor	237	2.63/.62	-.82	2.74/.56	-.18	10.92/3.89	2.38*
	Non-doctor	82	2.70/.66		2.76/.64		9.90/3.14	
G_Y (Friendly)	Doctor	231	2.69/.56	-1.79	2.81/.46	-.55	11.10/3.97	2.65**
	Non-doctor	82	2.80/.46		2.84/.46		9.95/3.13	

Note. # indicates the number of users in the group. * $p < .05$; ** $p < .01$.

According to the user perception questionnaire results, both the first-time and the feel-friendly users, both doctors and non-doctors, felt equally helpful in decreasing anxiety and increasing mastery. In other words, at the user perception level, both doctors and non-doctors believed that the HAI-based clinical safety training program could help them equally at the psychological level, such as decreasing anxiety and increasing mastering levels. Furthermore, the doctors spent more time answering questions, and answered more questions correctly while compared to the non-doctors. In response to RQ2 (exploring the impact of HAI -based training programs on the learning performance, mastery level, and user anxiety of different medical personnel in clinical workplace safety knowledge), the HAI-based clinical safety training program was more helpful to the doctors than the non-doctors. In addition to the help at psychological level, such as decreasing anxiety and increasing mastery levels, doctors also obtained a significantly higher learning performance than non-doctors.

To understand hospital personnel's points of view of the HAI-based clinical safety training program, we collected some opinions and suggestion from doctors and non-doctors as qualitative feedback. For example, Dr. Zhang mentioned, "This is a well-designed platform. For new doctors who are not familiar with the clinical environment, this simulation training can decrease their anxiety. It is a learning platform suitable for new doctors." Ms. Zhou, the chief nurse of Emergency Department had commented as "My work partners tell me that they are more likely to experience needle stick and sharps injuries or poor medical waste management during first aid. But with a simulation training system, this unnecessary risk of workplace hazards can be reduced, which give us more confidence in clinical practice." The responses of the above-mentioned hospital members were consistent with the questionnaire response results. Both doctors and non-doctors feel that the HAI-based clinical safety training program can help them decrease anxiety and increasing mastery level.

Yang et al. (2021) had mentioned that though AI had evolved rapidly and could somehow imitate human behaviors, the fundamental difference between artificial intelligence and human intelligence was emotion, feeling and cognition. To compensate the shortage that AI may make, this study, by embedding questionnaires into the system, showed the users experienced emotional well-being after using this learning system. Therefore, our study suggests that creating a user-friendly system is also effective for medical education.

By using real-time feedback with embedded questionnaire, our system is a user-friendly, humanity based, explainable and trustworthy education platform, i.e., a humanity-centered design. Shneiderman (2020) encourages researchers to strike a delicate balance between human control and computer automation, bring it a higher level of humanity and creativity to enhance HAI utilization. In this study, we evaluated learning performance, mastery level, and user anxiety, making it a human-centered design by further assessing psychological level. We also considered it is a good example of HAI because it recognizes human feeling and maintains adaptable automation, creating a trustworthy and explainable system.

5. Conclusions and limitations

After three years' consecutive study from April 2018 to December 2021, we found that the HAI-based clinical safety training system for the CWM and NSSI prevention could be applied to hospital-wide medical personnel. This study explored the user's learning performance and effectiveness including decrease anxiety and increase master level of clinical work safety knowledge after HAI-based training. From the first-time and feel-friendly user's experience, this training could achieve significantly higher learning performance, decreased anxiety, and increased mastery level of clinical safety knowledge. In comparison with non-doctor users, doctors gained more benefits, such as improved learning performance and effectiveness including decreased anxiety and increased clinical safety knowledge.

As for its limitation, the subjective values were collected via questionnaire which could be biased. Possible bias may also include missing or inadequate data for intended purpose, such as belief and behavior when being asked about hypothetical or personalized question. For instance, "Are you anxious about managing clinical wastes appropriately and needle stick/sharp injuries prevention?" "Are you familiar with clinical wastes disposal and needle stick/sharp injuries prevention after HAI-based learning?" or "Are you confident in clinical wastes disposal and needle stick/sharp injuries prevention?"

Finally, according to Accreditation Council for Graduate Medical Education (ACGME), there are 6 general competencies for residents as a milestone, namely of Patient Care, Medical Knowledge, Practice-based learning and improvement, Interpersonal communication skills, Professionalism, Systems-based practice (The Accreditation Council for Graduate Medical Education [ACGME], 1999). We expect to help medical students achieve these six competencies by HAI-based learning system in the future.

As for future prospective, after massive database collection, we hope to create a platform that could not only easily extend to clinical work safety, but also integrate with medical education or other issues. We are looking forward to building a platform to develop a user-friendly and customized learning program to all fields for different levels of doctors as well as different occupations.

In conclusion, the significant increasing of the learning performance, increasing mastery level and decreasing anxiety about knowledge and skills of the CWM and NSSI prevention indicates the high acceptability among users of Human-centered Artificial Intelligence courses. Therefore, it's necessary for educational committee to keep on selecting and establishing clinical safety related topics base on the initial positive findings and our HAI training system.

Acknowledgment

The authors were grateful to all the staff members and volunteer for their active cooperation during the whole study of this HAI-based simulation system. The reported research was funded by Taipei Veterans General Hospital [Grant number: 111EA-009, V111EA-010, V111C-018, V111C-038, VTA111-A-4-3], Ministry of education (PMN1100719), and Ministry of Science and Technology (Taiwan) [Grant number: MOST-110-2634-F-A49-005, MOST-109-2314-B-010-032-MY3 and MOST-110-2511-H-A491-504-MY3].

References

- Abebe, S., Raju, R., & Berhanu, G. (2017). Health care solid waste generation and its management in Hawassa Referral Hospital of Hawassa University, Southern, Ethiopia. *International Journal of Innovative Research & Development*, 6(5), 126-132.
- Accreditation Council for Graduate Medical Education. (1999). *1999 Annual report*. https://www.acgme.org/globalassets/PDFs/an_1999AnnRep.pdf
- Akkajit, P., Romin, H., & Assawadithalerd, M. (2020). Assessment of knowledge, attitude, and practice in respect of medical waste management among healthcare workers in clinics. *Journal of Environmental and Public Health*, 2020, 8745472.
- Ananth, A. P., Prashanthini, V., & Visvanathan, C. (2010). Healthcare waste management in Asia. *Waste Management*, 30(1), 154-161.
- Askarian, M., & Malekmakan, L. (2006). The Prevalence of needle stick injuries in medical, dental, nursing and midwifery students at the university teaching hospitals of Shiraz, Iran. *Indian journal of medical sciences*, 60(6), 227-232.
- Brockett, R. G., & Hiemstra, R. (2018). *Self-direction in adult learning: Perspectives on theory, research and practice*. Routledge.
- Buch, A. C., Rathod, H., & Naik, M. D. (2021). Scope and challenges of self-directed learning in undergraduate medical education: A Systematic review. *Journal of Medical Education*, 20(1), e114077. <http://doi.org/10.5812/jme.114077>
- Chartier, Y., Emmanuel, J., Pieper, U., Prüss, A., Rushbrook, P., Stringer, R., Townend, W., Wilburn, S., & Zghondi, R. (2014). *Safe management of wastes from health-care activities*. World Health Organization.
- Gao, X., Hu, B., Suo, Y., Lu, Q., Chen, B., Hou, T., Qin, J., Huang, W., & Zong, Z. (2017). A Large-scale survey on sharp injuries among hospital-based healthcare workers in China. *Scientific Reports*, 7(1), 1-7.
- Ghasemzadeh, I., Kazerooni, M., Davoodian, P., Hamed, Y., & Sadeghi, P. (2015). Sharp injuries among medical students. *Global Journal of Health Science*, 7(5), 320-325. <http://doi.org/10.5539/gjhs.v7n5p320>
- Gibelli, J., Aubin-Horth, N., & Dubois, F. (2019). Individual differences in anxiety are related to differences in learning performance and cognitive style. *Animal Behaviour*, 157, 121-128.
- Hayleeyesus, S. F., & Cherinete, W. (2016). Healthcare waste generation and management in public healthcare facilities in Adama, Ethiopia. *Journal of Health and Pollution*, 6(10), 64-73.
- Joshi, SC., Diwan, V., Tamhankar, A. J., Joshi, R., Harshada, S., Sharma, M., Pathak, A., Macaden, R., & Lundborg, C. S. (2015). Staff perception on biomedical or health care waste management: A Qualitative study in a rural tertiary care hospital in India. *PLoS One*, 10(5), e0128383. <https://doi.org/10.1371/journal.pone.0128383>
- Kalantary, R. R., Jamshidi, A., Mofrad, M. M. G., Jafari, A. J., Heidari, N., Fallahzadeh, S., Arani, M. H. & Torkashvand, J. (2021). Effect of COVID-19 pandemic on medical waste management: a case study. *Journal of Environmental Health Science and Engineering*, 19(1), 831-836.
- Khunger, N., & Kathuria, S. (2016). Mastering surgical skills through simulation-based learning: Practice makes one perfect. *Journal of cutaneous and aesthetic surgery*, 9(1), 27-31.
- Knowles, M. S. (1975). *Self-directed learning: A Guide for learners and teachers*. Association Press.
- Lin, T. J., & Lan, Y. J. (2015). Language learning in virtual reality environments: Past, present, and future. *Educational Technology and Society*, 18(4), 486-497.
- Letho, Z., Yangdon, T., Lhamo, C., Limbu, C. B., Yoezer, S., Jamtsho, T., Chhetri, P., & Tshering, D. (2021). Awareness and practice of medical waste management among healthcare providers in National Referral Hospital. *PLoS One*, 16(1), e0243817. <https://doi.org/10.1371/journal.pone.0243817>
- Maina, J. (2018). Knowledge, attitude and practice of staff on segregation of hospital waste: A Case study of a tertiary private hospital in Kenya. *European Scientific Journal*, 14(9), 401-417.
- Markovic-Denic, L. N., Mihajlovic, B., Cemerlic-Adic, N., Pavlovic, K., & Nicin, S. (2011). The Effect of training program to reduce needle stick injuries. In *BMC Proceedings*, 5(6), 1-1. BioMed Central. <https://doi.org/10.1186/1753-6561-5-S6-P217>
- Merandi, R., & Williams, A. (2017). Effectiveness of ‘training programme’ on knowledge and practices of biomedical waste management among health care workers. *Galore International Journal of Health Sciences and Research*, 2(4), 45-52.
- Norsayani, M. Y., & Hassim, I. N. (2003). Study on incidence of needle stick injury and factors associated with this problem among medical students. *Journal of Occupational Health*, 45(3), 172-178.

- Ozder, A., Teker, B., Eker, H. H., Altundis, S., Kocaakman, M., & Karabay, O. (2013). Medical waste management training for healthcare managers-a necessity? *Journal of Environmental Health Science and Engineering*, 11(1), 1-8. <https://doi.org/10.1186/2052-336X-11-20>
- Patterson, C., Crooks, D., & Lunyk-Child, O. (2002). A New perspective on competencies for self-directed learning. *Journal of Nursing Education*, 41(1), 25-31.
- Peng, J., Wu, X., Wang, R., Li, C., Zhang, Q., & Wei, D. (2020). Medical waste management practice during the 2019-2020 novel coronavirus pandemic: Experience in a general hospital. *American Journal of Infection Control*, 48(8), 918–921.
- Ricotta, D. N., Richards, J. B., Atkins, K. M., Hayes, M. M., McOwen, K., Soffler, M. I., Tibbles, C. D., Whelan, A. J., Schwartzstein, R. M., & Millennium Conference 2019 writing group. (2021). Self-directed learning in medical education: Training for a lifetime of discovery. *Teaching and Learning in Medicine*, 1-11. <https://doi.org/10.1080/10401334.2021.1938074>
- Robinson, J. D., & Persky, A. M. (2020). Developing self-directed learners. *American Journal of Pharmaceutical Education*, 84(3), 292-296.
- Rozinaj, G., Vančo, M., Vargic, R., Minárik, I., & Polakovič, A. (2018). Augmented/virtual reality as a tool of self-directed learning. In *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 1-5). IEEE. <http://doi.org/10.1109/IWSSIP.2018.8439309>
- Sharma, R., Rasania, S. K., Verma, A., & Singh, S. (2010). Study of prevalence and response to needle stick injuries among health care workers in a tertiary care hospital in Delhi, India. *Indian journal of community medicine: Official publication of Indian Association of Preventive & Social Medicine*, 35(1), 74-77. <http://doi.org/10.4103/0970-0218.62565>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504.
- Simpkin, A. L., & Walesby, K. E. (2017). Training tomorrow's doctors. *Future Hospital Journal*, 4(1), 56-60.
- Vellingiri, B., Jayaramayya, K., Iyer, M., Narayanasamy, A., Govindasamy, V., Giridharan, B., Ganesan, S., Venugopal, A., Venkatesan, D., Ganesan, H., Rajagopalan, K., Rahman, P. K. S. M., Cho, S.-C., Kumar, N. S., & Subramaniam, M. D. (2020). COVID-19: A Promising cure for the global panic. *Science of the Total Environment*, 725, 138277. <https://doi.org/10.1016/j.scitotenv.2020.138277>
- Wicker, S., Nürnberger, F., Schulze, J. B., & Rabenau, H. F. (2008). Needle stick injuries among German medical students: Time to take a different approach? *Medical Education*, 42(7), 742-745.
- Yang, J. C., Lin, M. Y. D., & Chen, S. Y. (2018). Effects of anxiety levels on learning performance and gaming performance in digital game-based learning. *Journal of Computer Assisted Learning*, 34(3), 324-334.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2021, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Yazie, T. D., Tebeje, M. G., & Chufa, K. A. (2019). Healthcare waste management current status and potential challenges in Ethiopia: A Systematic review. *BMC Research Notes*, 12(1), 1-7. <https://doi.org/10.1186/s13104-019-4316-y>

Artificial Intelligent Robots for Precision Education: A Topic Modeling-Based Bibliometric Analysis

Xieling Chen¹, Gary Cheng², Di Zou^{3*}, Baichang Zhong¹ and Haoran Xie⁴

¹School of Information Technology in Education, South China Normal University, Guangzhou, China //

²Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR //

³Department of English Language Education, The Education University of Hong Kong, Hong Kong SAR //

⁴Department of Computing and Decision Sciences, Lingnan University, Hong Kong SAR //

xielingchen0708@gmail.com // chengks@eduhk.hk // dizoudaisy@gmail.com //

zhongbc@163.com // hrxie2@gmail.com

*Corresponding author

ABSTRACT: As a human-friendly system, the artificial intelligence (AI) robot is one of the critical applications in promoting precision education. Alongside the call for humanity-oriented applications in education, AI robot-supported precision education has developed into an active field, with increasing literature available. This study aimed to comprehensively analyze directions taken in the past in this research field to interpret a roadmap for future work. By adopting structural topic modeling, the Mann-Kendall trend test, and keyword analysis, we investigated the research topics and their dynamics in the field based on literature collected from Web of Science and Scopus databases up to 2021. Results showed that AI robots and chatbots had been widely used in different subject areas (e.g., early education, STEM education, medical, nursing, and healthcare education, and language education) for promoting collaborative learning, mobile/game-based learning, distance learning, and affective learning. However, a limited practice in developing true human-centered AI (HCAI)-supported educational robots is available. To advance HCAI in education and its application in educational robots for precision education, we suggested involving humans in AI robot design, thinking of individual learners, testing, and understanding the learner–AI robot interaction, taking an HCAI multidisciplinary approach in robot system development, and providing sufficient technical support for instructors during robot implementation.

Keywords: Artificial intelligence robots, Topic modeling, Bibliometric analysis, Precision education, Research topics, Future of human-centered artificial intelligence

1. Introduction

Alongside the prevalence of artificial intelligence (AI) applications in personalized learning (e.g., robots and chatbots) is a shift from technology-driven to humanity-driven applications (Yang et al., 2021).

1.1. Human-centered AI and its use in education advancement

According to Yang et al. (2023), human-centered AI (HCAI) is interpreted as “AI taking humanities as the primary consideration, which requires explainable and trustworthy computation for continuously adjusting AI algorithms through human context and societal phenomena to augment human intelligence with machine intelligence, thereby enhancing the welfare of human kinds” (p. 1).

A robust, trustworthy HCAI system, when being applied in education, should have the capabilities of understanding individual learners’ prior experiences, needs, interests, relevant emotions, and social structures, adapting to complex real-world learning contexts, and appropriately interacting with individuals (Li et al., 2021). This is commonly achieved by allowing humans to seamlessly interact with and guide AI and enrich the AI system with human capabilities, knowledge about the world, and users’ personal perspectives (Renz & Vladova, 2021). HCAI also bridges the gap between machines and learners by leveraging emotional and cognitive input from learners and allowing machines to understand learners’ language, emotions, and behaviors (Shneiderman, 2020).

HCAI’s capabilities of understanding individuals, adapting to contexts, and appropriate interaction are particularly important in advancing education. This is because learning involves teaching and interaction with humans; thus, AI-supported learning technologies should be human-centered, focusing on both performance and learners’ emotions, feelings and outcomes, interaction, and learning contexts. (Shneiderman, 2020).

1.2. Precision education and HCAI

In HCAI, AI design is undergoing a transition from one-size-fits-all to precision approaches (Yang, 2021). As a core component of HCAI, precision education involves “the use of machine learning and learning analytics of AI to improve teaching quality and learning effectiveness” (Yang et al., 2021, p. 1-2) by “identify[ing] at-risk students as early as possible and provid[ing] them with timely intervention through [the four steps of] diagnosis, prediction, treatment, and prevention” (Yang, 2021, p. 106). For instance, in the case of poor performance and learning disabilities, learners’ learning behaviors, learning contexts, and learning strategies can be analyzed by following the four steps to identify solutions (Lu et al., 2018).

As precision education focuses on providing prevention and intervention to individuals, learning systems’ capabilities of integrating knowledgeable instructors’ expertise and intelligence into decision-making are essential (Hart, 2016). Developing such intelligent systems requires the ability to simulate educational experts’ intelligence (Hwang et al., 2020).

Currently, few AI systems for precision education have explicitly considered HCAI approaches. One mere example is provided by Weitekamp et al. (2020), in which instructors designed computerized lessons based on insights generated by an AI tutor. Although with limited human capabilities, such AI systems share similarities with HCAI design in caring for individuals’ needs and emotions, real-world contexts, and human-machine interactions (Renz & Vladova, 2021).

1.3. AI robots for precision education

AI robots or human-friendly systems are increasingly important for precision education by allowing personalized, natural interaction with real-life physical environments through practical demonstrations and hands-on experiences (Chen et al., 2020b). Practices regarding AI robots’ use in precision education are available. In Zhong et al. (2020), a quasi-experimental design was implemented with 84 junior high school students to show virtual and physical robots’ effectiveness in promoting students’ higher-order thinking in resolving complex problems and reducing cognitive load. In Santos et al. (2020), children shared their experienced emotional events to a chatbot, which then provided personalized scaffoldings accordingly.

Advantages of AI robots for precision education (Edwards et al., 2018) include (1) facilitating one-to-one learning by adapting instruction and communication to individual learners’ knowledge levels and learning styles, (2) promoting the shift of teachers’ roles toward overseers responsible for designing and selecting machine-oriented instruction, monitoring learner progress, and providing pastoral support, (3) turning abstract concepts into real-world problems adapted to individuals’ learning needs to promote all-in-one learning experiences where learners put theoretical knowledge into practice, and (4) supporting students to learn at their own pace with personalized materials through interactive experimental learning individually and collaboratively.

There are also disadvantages concerning AI robots’ use for precision education (Xia & Zhong, 2018; Tlili et al., 2020; Chen et al., 2022). First, it would be time-consuming and challenging to create, rebuild, and repair AI robots that are complex, with personalized learning objectives being considered. Second, robots with the same protocols for behavior analysis and pattern recognition and without the sense of humor or real-time life experiences cannot make personal connections to learners with things in life apart from the assigned work. Additionally, it is challenging to carry out robot cultivation on a large scale due to high requirements for cultivation resources and supporting facilities.

Despite the disadvantages, the available literature regarding AI robots used to promote precision education generally shows their positive effects on learning motivation, participation, and engagement, understanding of science processes and mathematical concepts, achievement score improvement, and development of creativity, designing, problem-solving, and teamwork (Tegos et al., 2011; She & Ren, 2021; Edwards et al., 2018; Kubilinskienė et al., 2017). Consequently, AI robot-supported precision education has developed into an active field of research.

1.4. Reviews on HCAI and educational robots

Discussion on HCAI in education is available. Renz and Vladova (2021) demonstrated the need for HCAI practice to promote the human condition for precision and smart learning. However, there is currently no literature review on HCAI in education due to very limited pedagogical practices. Regarding AI robot-supported

precision education, based on a review of 22 empirical studies concerning robotics education in K-12, Xia and Zhong (2018) identified the prevalence of LEGO robots and non-experimental design and highlighted instructional suggestions about open environments, targeted design, appropriate pedagogy, and timely support. Guided by activity theory, Tlili et al. (2020) analyzed 30 studies about robot-supported special education, identifying research gaps, challenges, and contradictions. These reviews have advanced knowledge about educational robots; however, they mainly adopted systematic review methodologies that are prone to error and coding inconsistencies. Additionally, considering the increased research on AI robots' applications in education, a comprehensive examination is needed to understand the directions taken in the past to enlighten a roadmap for future work.

1.5. Research aims and questions

This study analyzes extant literature regarding AI robot-supported precision education using topic modeling and keyword analysis. We mainly focus on research topics and their dynamics by looking at research topics' evolution across four chronological sub-periods of time during the past 20 years and relating them to technological, pedagogical, and methodological advances. Given the low number of papers in the early years, according to López-Robles et al. (2019), a suitable way is to divide the time span into comparable periods. For example, López-Robles et al. (2019) divided the study period (1988–2017) into 1988–1997, 1998–2007, 2008–2012, and 2013–2017 with 144, 970, 2,083, and 3,195 papers, respectively. In Cobo et al. (2011), five non-equidistant periods of time (i.e., 1978–1989, 1990–1994, 1995–1999, 2000–2004, and 2005–2009) were used because there were few papers in the former years, which could lead to a low number of keywords being used as input in co-words analysis to detect main themes. Accordingly, this study uses non-equidistant periods of time (i.e., 2001–2010, 2011–2017, and 2018–2019, with 26, 40, and 49 papers, respectively) to ensure a good input for data analysis. We additionally include the sub-period 2020–2021 to understand the most recent research topics in the field. Findings can help educators understand AI robots' potential to promote precision education. Based on the updated information on the present research, we provide insights into future research and pedagogical practices in this field. There are four research questions (RQs).

RQ1: What were the major research topics during 2001–2010?

RQ2: What were the major research topics during 2011–2017?

RQ3: What were the major research topics during 2018–2019?

RQ4: What were the major research topics during 2020–2021?

2. Methodology

2.1. Data collection

On 24 October 2021, Web of Science and Scopus databases were searched to identify articles about AI robot-supported precision education. Figure 1 shows the flowchart of data collection. There were two search strategies. The first strategy involved search terms concerning AI, robots, personalization, and education. The search terms were decided with reference to Chen et al. (2021) and Zawacki-Richter et al. (2019) by considering both personalized learning and the use of AI robots. The second strategy considered personalization-, education-, and chatbot-related terms determined by referring to Chen et al. (2021) and Smutny and Schreiberova (2020) by considering both personalized learning and the use of chatbots. Data were limited to journal articles or conference papers written in English.

After duplication, 5,112 papers were included for screening based on exclusion criteria presented in Figure 1. When deciding to include a paper or not, we started from the first criterion and directly excluded 4,784 papers irrelevant to instruction and learning. Subsequently, for the remaining 328 papers, we checked whether they provided detailed information about robots' use for educational purposes. Accordingly, 73 papers were excluded. We further excluded 81 reviews or survey papers. Finally, 174 papers remained for data analysis. Figure 2 shows the number of papers by year, which indicates two stages of development, that is, a slow-growth trend from 2001 to 2017 and a fast-growth tendency since 2018.

2.2. Data analysis

We analyzed the 174 papers using keyword analysis, structural topic models (Roberts et al., 2019), and the Mann-Kendall trend test. For keyword analysis, we extracted key phrases from titles and abstracts and calculated their frequencies in four sub-periods of time. For topic modeling, we first collected terms from titles and abstracts, then used term frequency-inverse document frequencies to filter unimportant terms. Exclusivity and semantic coherence criteria were used to facilitate model selection (Figure 3). A manual comparison of models with 13 and 15 topics based on associated papers and terms was further conducted, which indicates that the model with 13 topics produces “the greatest semantic consistency within topics and exclusivity between topics” (Chen et al., 2020a, p. 4). Two experts then examined the statistical results to determine topic labels. Among the 13 topics, one was excluded as it mixes up robot motion and learner motivation. The remaining 12 topics were included for analysis, with their changes in prevalence being examined using a trend test.

Figure 1. Flowchart of data collection

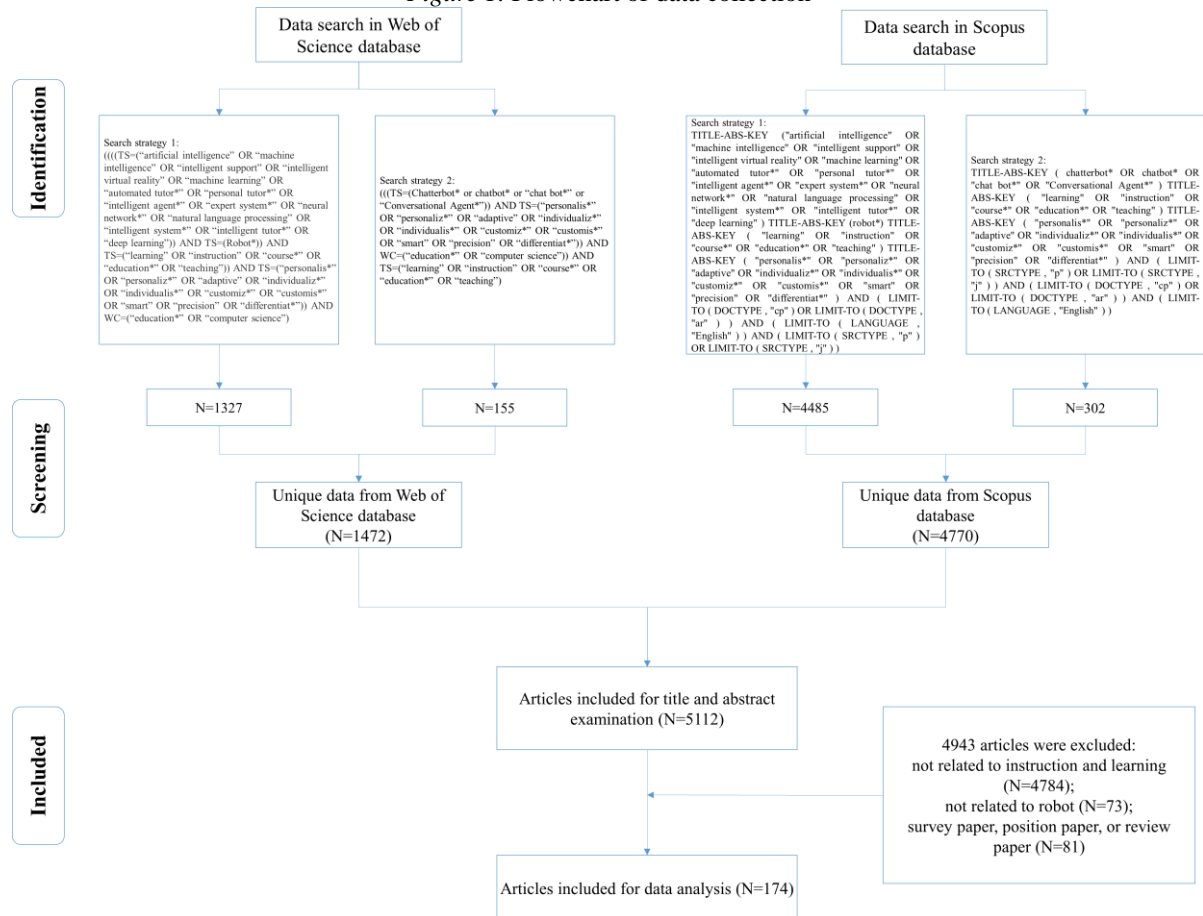


Figure 2. Number of papers by year

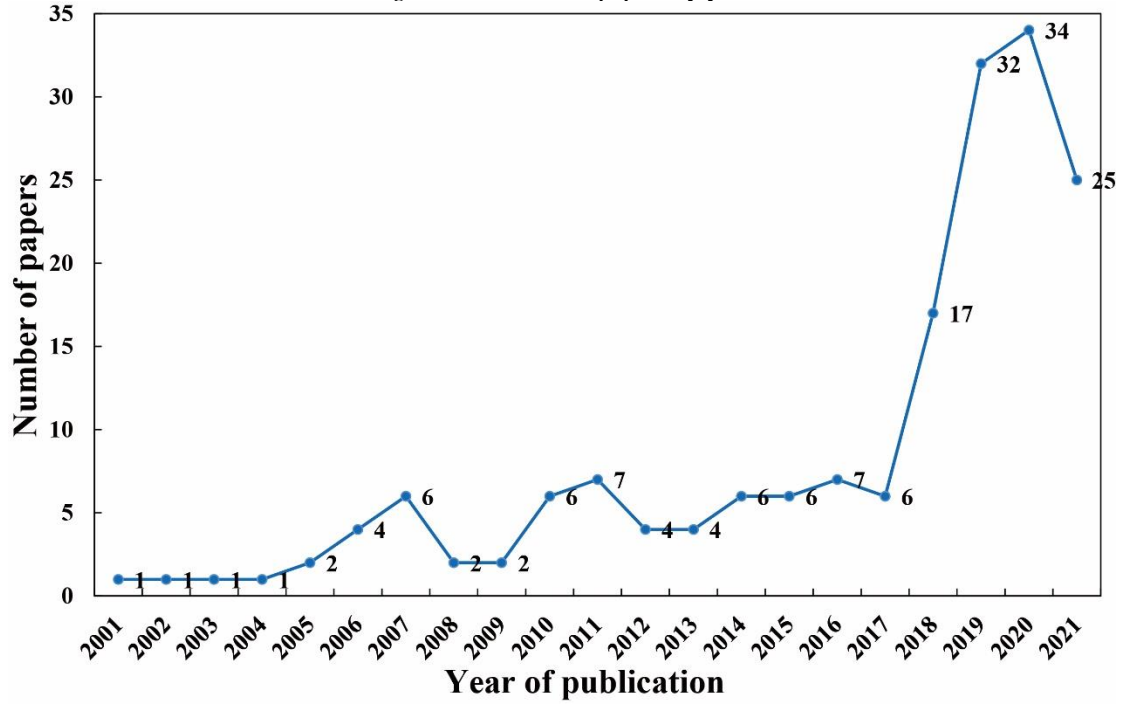
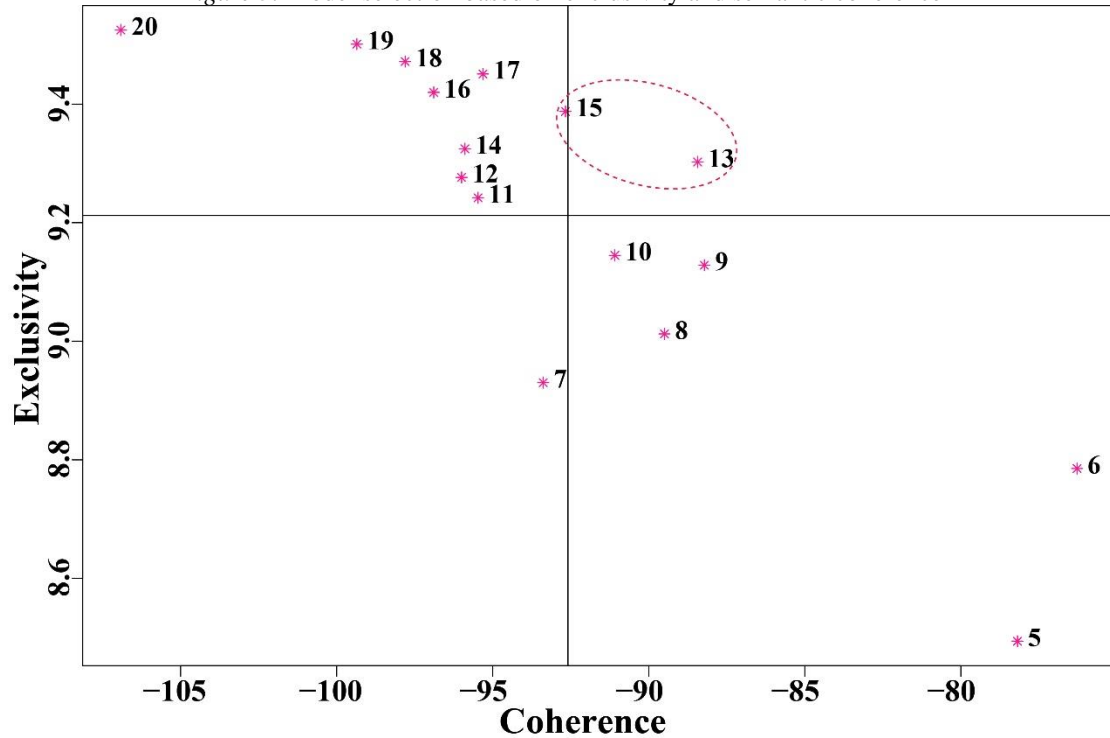


Figure 3. Model selection based on exclusivity and semantic coherence



3. Results

Figure 4 visualizes key phrases in the four sub-periods. Figure 5 visualizes emerging phrases during 2020–2021. During 2001–2010, researchers mainly focused on the conversational agent’s pedagogical use, especially in language learning (see Figure 4(a)). Subsequently, AI, collaborative learning, learning style, serious games, human-robot interaction, learning process, and adaptive learning received a growth of research interest among scholars (see Figure 4(b)). During 2018–2019, there was a growing research interest in neural networks, social robots, educational chatbots, young children, and machine learning (ML) (see Figure 4(c)). During 2020–2021, there was a trend in research on humanoid robots, natural language processing (NLP), deep learning, speech

recognition, argumentation skill, mental model, emotion recognition, dialog-based form, emotional engagement, adaptive writing support system, and cognitive load (see Figure 4(d) and Figure 5). Figure 6 presents the topic modeling results, and Figure 7 visualizes topic proportions by year, which clearly shows how the prevalence of each topic changed with time going on. The two most popular topics were *conversational agents* and *chatbots for education*, and four topics were increasingly researched, including *robots for early childhood education*, *chatbots for education*, *robots for STEM education*, and *chatbots for distance learning*.

Figure 4. Key phrases during sub-periods (a) 2001–2010, (b) 2011–2017, (c) 2018–2019, and (d) 2020–2021

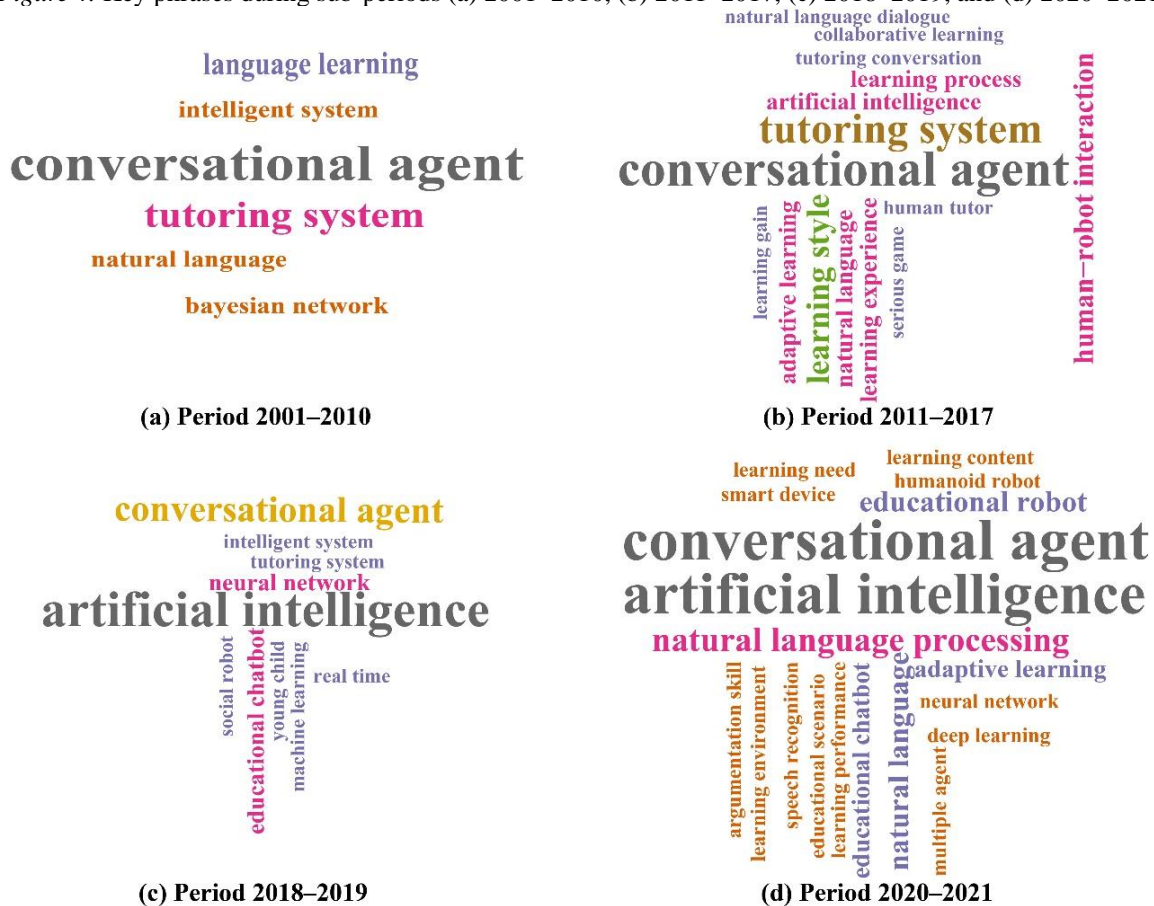


Figure 5. Emerging phrases during 2020–2021



Figure 6. Topics with suggested labels, topic proportions, and developmental trends

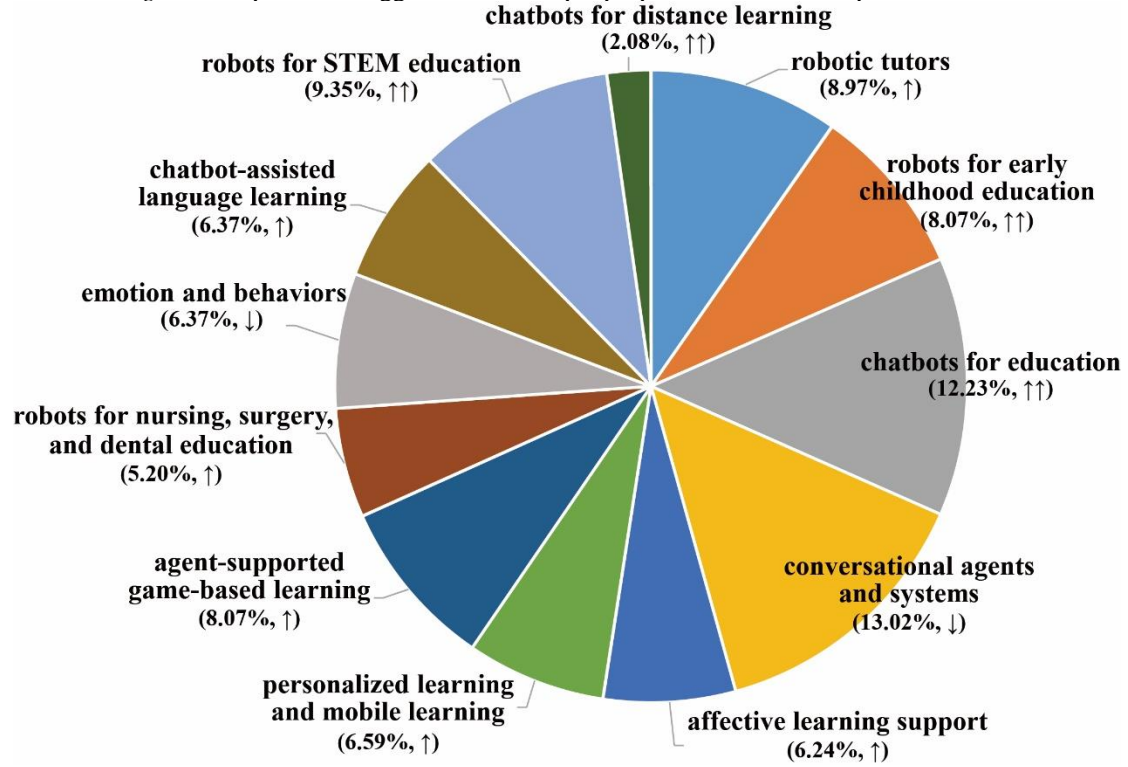
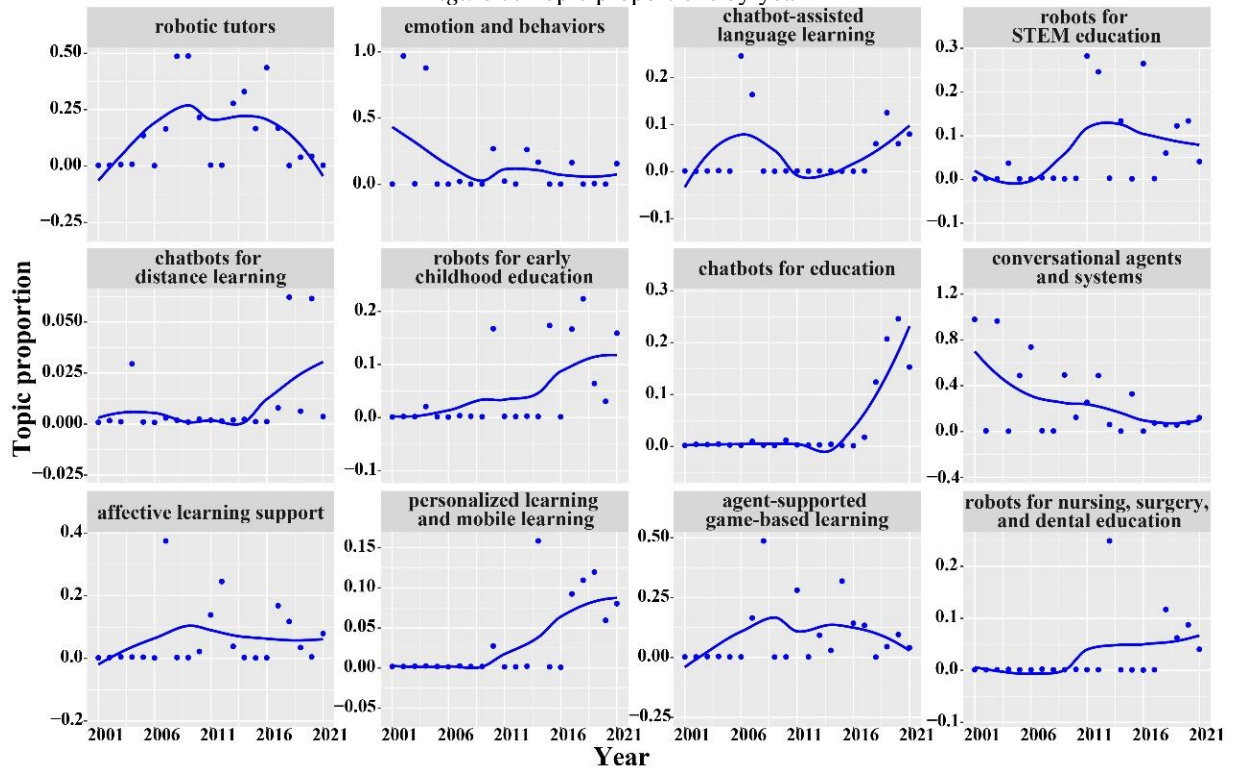


Figure 7. Topic proportions by year



4. Discussion

This study provides a topic modeling-based bibliometric analysis of literature related to AI-supported robots for precision education to understand the most frequently studied topics in the field during the past 20 years. Results showed rapid growth of interest in AI robot-supported precision education research because of advances in computers, information communication technologies, and analytical innovations like AI and ML, alongside

educators' increased interest in exploring AI robots' potential for personalized education (Chen et al., 2021). In line with the four RQs, the following sub-sections present a discussion on the findings of research topics in the four sub-periods. We further discussed the challenges and directions to advance the development of HCAI and its application in educational robots.

4.1. Research topics during 2001–2010

In this period, the main topics centered on conversational agents for educational use and embodied conversational agents for affective learning, evidenced by the high frequencies of phrases such as conversational agent, language learning, intelligent/tutoring system, and the topic of *emotion and behaviors*, as indicated in Figure 4(a) and Figure 7.

4.1.1. Conversational agents for educational use

Conversational agents, which allow humans to interact with computer systems using natural language socially and effectively, endorse sociocultural theory's emphasis on learning through social participation and interaction (Vygotsky, 1978). This is particularly important in language learning, where learners need direct and frequent social interaction for target language practice in authentic exchanges, accordingly to language socialization and situated language learning. Intelligent tutoring systems integrated with conversational agents, by extending "communication and interactive opportunities beyond [the] lecture experience" (Gosper et al., 2008, p. 1), offer enormous opportunities to cultivate learners' spontaneous productive skills and second language fluency. Alongside the advances in AI, and especially NLP and speech recognition technologies, conversational agents' use for supporting social learning has become affordable. For example, CALMsystem (Kerly et al., 2008) supported a learner's reflection by inferring a knowledge level for the learner depending upon his answers and encouraging him to involve in a dialogue to reflect on his performance.

4.1.2. Embodied conversational agents for affective learning

Alongside conversational agents' prevalence in education is an attempt to exploit their emotional capabilities to deal with learners' affects productively. The correct identification of learners' emotions is essential to learning endeavors and outcomes because emotions expressed during social interaction can affect attention, meaning creation, and memory, thus influencing learners' cognitive and affective development. This task can be achieved by using embodied conversational agents, which, with a virtual animated body that produces both verbal and non-verbal signals, can show empathy and emotions and support learners' emotional states productively. According to De Waal (2009), conversational agents programmed with natural language that includes emotions and empathy promote stronger relationships and collaboration and more complex learner–conversational agent interactions. Increasingly, embodied conversational agent is employed as an interaction metaphor in education. Morton and Jack (2005) integrated speech recognition with embodied conversational agents and virtual worlds to construct immersive, contextualized environments where learners conversed in the target language and obtained feedback from embodied conversational agents.

4.2. Research topics during 2011–2017

In this period, issues related to embodied conversational agents' integration into digital games, conversational agents for computer-supported collaborative learning, AI robots in medical, nursing, and healthcare education, and AI robot-supported STEM education became popular, with phrases such as collaborative learning, serious games, and topics of *agent-supported game-based learning*, *robots for STEM education*, and *robots for nursing, surgery, and dental education* appearing increasingly, as shown in Figure 4(b) and Figure 7.

4.2.1. Embodied conversational agents' integration into digital games

The increasing interest in integrating embodied conversational agents into digital games is driven by the need to make digital game-based learning more interactive. Embodied conversational agents and digital games have close relationships. First, interactivity and believability are salient characteristics in digital game-based learning to fully engage learners, which are affordable by embodied conversational agents. Second, gamification's ability to appreciate users' motives, cognition, and emotions to optimize their feelings, motivations, and engagement

corroborates embodied conversational agents' capabilities to provide affective and emotional support. Furthermore, both digital games and conversational agents promote social skill development. Additionally, embodied conversational agents enrich learning experiences in gamification through active experimentation and multimodal interaction (e.g., gaze, facial expressions, and gestures), thus making learning more experiential (Colpaert, 2006). Consequently, embodied conversational agents are increasingly integrated into digital games to enrich verbal and non-verbal interaction, especially for social communication skill promotion. For example, a serious game, ECHOES, adopted an embodied conversational agent as autistic children's social companion to help them develop social communication skills (Bernardini et al., 2014).

4.2.2. Conversational agents for computer-supported collaborative learning

Alongside the call for socio-cognitive learning (Vygotsky, 1978) that emphasizes learning through socialization and collaboration is the increasing use of conversational agents as personalized tutoring aids to promote computer-supported collaborative learning. In computer-supported collaborative learning, a temporary appearance of a suitable degree of misunderstanding is beneficial; thus, a supportive conversational agent should intelligibly elicit peer dialogue to foster learning beneficial conditions during collaboration. An agent represented as a three- or two-dimensional human-like avatar or interface in computer-supported collaborative learning environments can (a) trace learning processes, (b) stimulate interaction and collaboration, and (c) inform learners about interaction status. For example, a conversational agent in a web collaborative learning system (Tegos et al., 2011) intelligently facilitated and triggered discussion among partners by allowing instructors to define agent interventions when an important concept was detected in learners' dialogue.

4.2.3. AI robot-supported medical, nursing, and healthcare education

In medical and healthcare education, simulated training provides valuable opportunities for students to acquire required skills and rehearse skills learned for future careers. In simulated training, high-fidelity simulators are a necessity. There are two common types of simulated patients as the recipient of students' skills, including stationary manikins and human simulated patients. However, stationary manikins cannot reproduce human movements or respond to trainees' commands, and human simulated patients have difficulties in exactly imitating real patients, which usually leads to ineffective and inefficient simulations. To provide effective simulated training, educators increasingly exploit robots' ability to simulate required actions in supporting medical and clinical simulated training.

Due to the shortage of qualified nurses while the ever-aging population, simulator robots are increasingly popular in nursing training, particularly regarding the patient transfer, to simulate patient's limb movements to help nursing students learn nursing skills. Researchers also exploit robots' ability to express emotions and feel pain like humans via visual-based feedback. In Lee et al. (2021), a robot's pain level was calculated using fuzzy logic and displayed in real time by a projector and a three-dimensional facial mask during nursing training. Regarding emotion expression, by exploiting embodied conversational agents' capability of engaging in natural interaction with humans through dialog and non-verbal expressions, Bickmore and Gruber (2010) used embodied conversational agents as virtual counselors to offer problem-solving skill training and emotional support for caregivers.

4.2.4. AI robot-supported STEM education

In an increasingly complicated world, it is essential for youth to foster contextualized knowledge and skills covered by STEM to resolve complex problems and make sense of information. With the advances in robotics and automation, robots have become accessible for school-age children to facilitate their STEM learning by allowing them to explore their ideas using technical- and computational-oriented tangible objects. Robots' effectiveness in STEM education corroborates the idea of "making", which, rooted in constructionism, is increasingly brought into classrooms to engage children in various technology-enhanced making activities like building robotics inventions. Making with robotics is student-centered with a focus on constructionist learning, where "students engage in manipulating, assembling, and reassembling materials while going through the design learning process and problem-solving program errors through trial and error" (Eguchi, 2017, p.16). Such experience promotes transdisciplinary learning where learners encounter different concepts in STEM contextually; in this way, abstract concepts become visible and tangible for learners to comprehend when they test their ideas with robotics inventions.

4.3. Research topics during 2018–2019

In this period, the major topics included robots' use in early education, AI robots' integration into mobile learning, and neural network-based educational robots, witnessed by the increasing use of phrases such as neural networks, social robots, young children, and topics of *personalized learning and mobile learning* and *robots for early childhood education*, as indicated in Figure 4(c) and Figure 7.

4.3.1. Robots' use in early education

Prior to this period, robots' pedagogical affordances were mostly demonstrated in primary, secondary, and higher education, whereas this period has witnessed considerable interest in robots' use in early education. This is driven by the need to cultivate technology and innovation literacies at an early age. However, the traditional early childhood curriculum pays little attention to developing early knowledge about the artificial world. There is thus a call for systematic educational reform by encompassing technology with creative thinking and problem-solving in early childhood education to prepare children as future citizens in a fundamentally technology-driven society. Alongside advances in novel interfaces, programming languages, and robotics engineering, educational robotics kits that developmentally fit young children are increasingly available for them to engage in "learning by designing" and "learning by programming" activities through hands-on experiences. Fachada (2018) confirmed smart toy robots' effectiveness in promoting children's social engagement and conversation skills. Williams et al. (2019) highlighted that allowing young children to construct, program, test, and interact with their social robots through hands-on experiences promoted their understanding of how AI works.

Affordances of robots in early education included: (1) enabling young children to understand things they meet in daily life through playful, practical hands-on activities, (2) exploiting computational thinking-focused activities to facilitate active learning, enhance motivation, and maintain engagement, and (3) serving as emotionally learning companions to promote their social and language skills. The last affordance is especially important in the education of children with special needs, a field that has gained increased attention as our society aims to provide equal opportunities to these children to develop skills and improve their quality of life (Moyi, 2019). Particularly, intense concern has been attached to autistic children's education using intelligent robots. Most autistic children have difficulties socializing with others, but they have no problem communicating with objects like robots that offer human-like social cues, which, together with the simplicity of an object, can facilitate their social skill learning.

4.3.2. AI robots' integration into mobile learning

Alongside the global trend in ubiquitous mobile learning and the need for ever-present hands-on learning opportunities (Axelsson et al., 2019), educators attempt to involve learners in learning by developing robot-supported pedagogical models around mobile devices. Mobile robots have been popular in remote laboratories to allow students to explore and interact with the real world through sensors and actuators to learn a wide range of knowledge in programming, electronics, and robotics and enable resource sharing without time and space constraints.

Researchers have also coupled multimodal conversational interfaces to improve mobile applications with intelligent, communicative capabilities and adaptation to learners' requirements by enabling learners to interact directly with mobile conversational agents to accomplish tasks. In Kim et al. (2019), students conversed with and answered questions raised by a mobile chatbot via text to practice English grammar skills.

4.3.3. Neural network-based educational robots

The increase in adopting advanced neural network-based algorithms is driven by the need for "smart services" with cognitive and intellectual abilities that are more scalable to satisfy personalized learning needs. For example, in an intelligent learning assistant for autistic children (Vijayan et al., 2018), a deep conventional neural network model processed brain image patterns to make predictions about children's behaviors, a recurrent conventional neural network analyzed facial images for decision making, and a reinforcement learning module analyzed children's speech to make responses accordingly. In an educational chatbot developed by Sreelakshmi et al. (2019), a question-answering module used neural networks to extract suitable answers from the knowledge base, and a quiz generation module identified key sentences and generated question-answer pairs to generate quizzes for learners.

4.4. Research topics during 2020–2021

In this period, the major topics included chatbots' use in distance education, AI robots for argumentation skill acquisition, and integration of physiological sensors and advanced deep learning into educational robots, witnessed by the increased use of phrases such as deep learning, dialog-based form, argumentation skill, adaptive writing support system, mental model, emotional engagement, and cognitive load, and topics of *chatbot-assisted language learning*, *chatbot for distance learning*, and *chatbot for education*, as indicated in Figure 4(d), Figure 5, and Figure 7.

4.4.1. Chatbots in distance education

In the era of "Education 4.0", the call for integrating innovative AI technologies into blended and flipped classrooms results in the proliferation of distance education. Distance education is the fastest-growing educational modality driven by the wide affordances of digital and handheld devices and global Internet access. However, online learning is criticized for lacking support, feedback, and interaction and causing learners' sense of isolation. These limitations became apparent during the COVID-19 pandemic when there was a rash transition from traditional face-to-face classes to complete online education, thus urging educators to effectively tackle the limitations. Chatbots appear as an alternative to this impasse, as they can minimize manual effort and provide immediate user-friendly assistance, human-like interaction, and continuous psychological and pedagogical support anytime and anywhere. Consequently, scholars are attempting to integrate chatbots or virtual assistants into distance education platforms to enable greater interactivity, facilitate sociability, and make online learning more interactive and dynamic. In a personalized dialogue-based system (Rajkumar & Ganapathy, 2020), a chatbot scaffolded learners' learning by answering frequently-asked questions, recommending tutorials, and planning learning paths. In Seering et al. (2020), a social chatbot in online communities "grew up" from "birth" through its teenage years, interacting with community members and "learning" vocabularies from their conversations. By taking the personalization and interaction levels to a new height, chatbots ultimately reduce dropout rates and increase educational achievements and satisfaction among distance learners.

4.4.2. AI robots for argumentation skill acquisition

Compared to previous periods when chatbots in language education mostly centered on basic conversation and language skill development, in this period, there is an increase in using robots to facilitate the development of metacognition skills such as arguing in a reflective and well-formed manner, which are beneficial to cultivate communication, collaboration and problem-solving competencies (Wambsganss & Rietsche, 2019). To cultivate such skills as argumentation, individuals need to receive constant tutoring and feedback during learning. However, it is hard for instructors to offer adaptive support and feedback to individual learners, particularly in large-scale lectures or distance education. The recent advances in NLP and ML promote new pedagogical human-computer interaction by implementing adaptive personal computer assistants with argumentation mining approaches to access individuals' argumentation levels and provide adaptive feedback and step-by-step guidance to intelligently support argumentation learning, thus enabling individuals to learn autonomously and independently of instructors, time, and place.

4.4.3. Integration of physiological sensors and advanced deep learning into educational robots

Driven by the rise of Robotics 4.0 with prevalent disruptive technologies like the internet of robots, AI of Things, and deep learning, there is a trend in integrating physiological sensors like eye-tracking and advanced deep learning into educational robots. For example, eye-tracking signals can be collected from learners during their interaction with educational robots to understand changes in their workload, dynamics of emotions, and physiological state. In a Dinus intelligent chatbot (Majid & Santoso, 2021), sentiment analysis was adopted to identify learner emotions in textual-based conversation, and recurrent neural networks were used to classify the emotions based on current conversations. In a robot system for supporting autistic children (She & Ren, 2021), a neural network as a generative conversational agent generated meaningful and coherent dialogue responses, and a transfer learning module learned dialogue characteristics to resolve the limitation of insufficient dialogue corpus.

4.5. Challenges regarding AI robots' application in education

Regarding challenges about AI robots' application in education, researchers have noted instructors' acceptance of AI robots and technological challenges.

4.5.1. Instructors' acceptance of AI robots

Instructors are the keystone to AI robots' pedagogical implementation. Although many instructors could appreciate robotics' benefits, they are reluctant to use it. This is especially true for instructors who lack experience with information technologies and struggle to execute effectively on-the-spot responses to analytics from AI systems, thereby hindering robotics' application in education. To promote AI robots' pedagogical practice, it is essential to improve instructors' acceptance by showing them AI robots' pedagogical benefits via longitudinal experiments based on educational theories. Currently, the advantages of most AI-supported educational robots commonly exist in theory without evidence showing their effectiveness in real-world teaching and learning. This is because the experimental design for AI system assessment is challenging as large samples are needed to produce probabilistic results (Chen et al., 2021). However, such experiments should be promoted to improve instructors' acceptance and verify AI robots' true effectiveness in the long term rather than due to novelty effects. By putting theoretical advantages demonstrated in literature into practice, the pedagogical applicability of AI robots to realistic educational scenarios can be evaluated.

4.5.2. Technological challenges

One advantage of AI robots is the rich interaction with real-life environments. Although physical exercises or objects can be exploited to facilitate instruction as robots are physically present, currently, motor activities with robots are rarely integrated into learning tasks owing to their feasibility. This is because the more robots act and move through space, the more likely they are to induce technical issues, e.g., falling over or overheating. Nevertheless, as robot technologies advance, it is promising that motor activities would become feasible to integrate to trigger higher learning gains.

Although AI robots are intensively applied to facilitate language learning, their technological capabilities have limitations centering on inappropriate interpretation and response (e.g., failed communication when learners input incomplete sentences and chatbots respond with nonsense outputs and diminished learning interest due to limited emotion and visible cues) (Huang et al., 2022). To make robots autonomous in natural interaction, there is a need for effective action selection based on the understanding of learners' abilities and progress to trigger appropriate actions to scaffold their learning. Although many AI robots allow learners to learn conveniently by self-deciding what and how to learn, troubles arise when they cannot handle learning tasks or use robots appropriately. Thus, instructors need to monitor learners throughout their interactions with AI robots and provide scaffoldings when necessary.

4.6. Future of HCAI in education and its application in educational robots

To address the technological challenges of AI robots to promote a higher level of personalization and enhance instructors' and learners' acceptance of AI robots, there is a need to consider higher-level HCAI capabilities in educational robots.

4.6.1. Benefits of HCAI systems compared to traditional AI systems

Compared to traditional AI systems with difficulties in guaranteeing non-discrimination, due process, and understandability in decision-making, HCAI systems have unique benefits of informed decision-making, reliability and scalability, personalized learner experiences, and more inclusive outcomes (Shneiderman, 2020). First, by leveraging the power of humans and machines, HCAI contributes to more precise AI algorithms built from human input and values, thus enabling instructors to make highly informed decisions and design more adaptive support to promote students' better learning. Second, by exploiting technology's computational abilities and simultaneously leveraging emotional and cognitive inputs from humans, HCAI contributes to expanding processes and information to a larger volume without threatening data integrity or increasing human resource costs. Third, by considering learners' characteristics, needs, and learning behaviors during AI system development, HCAI contributes to personalized, fulfilling learner experiences. Additionally, by keeping humans

in the loop while building AI, HCAI enables humans to monitor for bias in algorithmic decisions, thus contributing to checked and balanced systems that make outcomes more inclusive.

However, HCAI's benefits can be constrained due to a high requirement of expertise and a lack of holistic assessment of HCAI approaches (Xu et al., 2022). As our results showed, although AI robots and chatbots have been widely used in different subject areas (e.g., early education, STEM education, nursing education, and language education) for promoting computer-supported collaborative learning, mobile/game-based learning, distance learning, and affective learning, limited practice on developing true HCAI-empowered educational robots is available. The limited practice of HCAI for educational purposes is also indicated in previous studies (e.g., Renz & Vladova, 2021).

4.6.2. Future of HCAI and its application in educational robots

To advance HCAI specifically to the community of AI robot-supported precision education, the concept of “co-learning” is important, which focuses on humans’ interaction with, learning from/with, and growing with AI (Huang et al., 2019). Specifically, AI needs to learn how to explain the learning, reasoning, and planning process to humans; humans need to learn how to include human intention and values in AI, explore ways to seamlessly interact with and teach AI, and adapt rules to enrich AI with uniquely human capabilities, knowledge about the world, and specific user’s personal perspective (Stephanidis et al., 2019). Future efforts on developing true HCAI systems for educational purposes are listed below.

Humans as part of a continuous feedback loop with AI. Being involved in the training, testing, and tuning processes of AI model construction, humans can validate AI decisions’ precision and offer feedback to AI in case of a wrong decision (Nakao et al., 2022). An example is given by Weitekamp et al. (2020), who allowed an instructor to teach an intelligent tutor who then taught learners. Specifically, a human instructor demonstrated to the tutor how to resolve problems. When the tutor provided wrong solutions, it showed to the human instructor learners’ trouble spots as ML systems usually encountered similar problems as learners.

Think of individual learners. To ensure that the end result enhances and positively augments the learning of individual learners, there is a need to clearly understand their backgrounds, needs, locations, and the ways they are going to utilize AI systems (Xu et al., 2022). This can be achieved by involving a sample of end learners in the model training, validating, and testing during system construction to capture their feedback.

Test and understand learner–AI robot interaction. Understanding and testing learner–AI robot interaction in real-world situations is essential for successful learner experiences (Xu et al., 2022). To promote AI robots’ capabilities in perceiving and interpreting complex real-world environments, human actions, and interactions (Li et al., 2021), there is a need to include more human-like world understanding and common-sense knowledge grounded in physical reality into AI robots by leveraging social and cultural theories (e.g., activity theory and actor-network theory) to frame the relationships between AI robotics and learning (Oliver, 2011) into social and cultural contexts. These frameworks help to see how learners make personal, social, and cultural meanings from interaction with robots, instructors, and peers to understand their learning trajectories and make sense of their learning experiences. The process of tracing learners’ interactions also helps understand how AI robots can be associated with their perspectives, interests, needs, and situated contexts to inspire feasible pedagogical implications for personalized instruction.

Take an HCAI multidisciplinary approach. As a multidiscipline, the successful HCAI design for learning objectives requires close collaboration among AI engineers and professionals, educational experts, psychologists, designers, sociologists, etc. to consider pedagogical innovations and learners’ learning styles, analyze learners’ behaviors during their interaction with AI robots in different contexts, and build technologically and pedagogically sound HCAI robots (Chen et al., 2021). For instance, for the development of an interactive intelligent tutoring system, which presents results related to classification, clustering, and prediction to learners via learner interfaces, educational experts can support the mental modeling of target learners (Xu et al., 2022).

Sufficient technical support. Training programs that emphasize well-balanced interactions among knowledge of contents, pedagogies, and technologies, can be provided to instructors with varied linguistic, instructional, and technological skills to guide them in effectively integrating AI robots into classrooms. During instruction, according to Bers et al. (2014), every instructor ought to have trained assistants to support troubleshooting technology issues, tracking children’s progress, and offering one-on-one help to achieve the optimal combination of human and AI robot instruction to best support students’ learning.

4.7. Limitations of this study

This study has limitations. Firstly, our analysis was based on records retrieved from Web of Science and Scopus. Although Web of Science and Scopus are multidisciplinary databases of academic output and are commonly adopted for literature reviews, there might still be articles related to AI robot-supported precision education that were not included in the two databases. Future research may consider exploring how research trends in AI robot-supported precision education vary when including articles from more journal sources and even from relevant conference proceedings. Furthermore, although topic models are acknowledged for their abilities to uncover thematic structure within large-scale literature data, they might not bring about strict conclusions. Future work can be conducted to combine text mining technologies with systematic and qualitative analysis methodologies to achieve a more fine-grained understanding. This would require developing techniques that allow systematic analysis of a large dataset to be conducted in an automatic way.

5. Conclusion

This study examined research on AI robots for precision education during 2001–2021 using structural topic modeling and keyword analysis. Results showed that AI robots and chatbots are widely used in different subject areas for promoting computer-supported collaborative learning, mobile/game-based learning, and affective learning. We also identified a lack of practice and research on true HCAI educational systems. Findings obtained contribute to identifying the main topics and gaps in the extant literature with implications for future practice and research on AI robots. Based on the findings, we propose suggestions for advancing HCAI and its application in educational robots from the perspective of “co-learning” between humans and AI. These suggestions include (1) involving humans as part of a continuous feedback loop with the AI model, (2) thinking of individual learners, (3) testing and understanding learner–AI robot interaction, (4) taking an HCAI multidisciplinary approach in system development, (5) providing sufficient technical support for instructors during AI robots’ implementation for educational purposes.

References

- Axelsson, M., Racca, M., Weir, D., & Kyrki, V. (2019). A Participatory design process of a robotic tutor of assistive sign language for children with autism. In *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1–8). <https://doi.org/10.1109/RO-MAN46459.2019.8956309>
- Bers, M. U., Flannery, L., Kazakoff, E. R., & Sullivan, A. (2014). Computational thinking and tinkering: Exploration of an early childhood robotics curriculum. *Computers & Education*, 72, 145–157. <https://doi.org/10.1016/j.compedu.2013.10.020>
- Bernardini, S., Porayska-Pomsta, K., & Smith, T. J. (2014). ECHOES: An Intelligent serious game for fostering social communication in children with autism. *Information Sciences*, 264, 41–60. <https://doi.org/10.1016/j.ins.2013.10.027>
- Bickmore, T., & Gruber, A. (2010). Relational agents in clinical psychiatry. *Harvard Review of Psychiatry*, 18(2), 119–130. <https://doi.org/10.3109/10673221003707538>
- Chen, X., Zou, D., Cheng, G., & Xie, H. (2020a). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A Retrospective of all volumes of *Computers & Education*. *Computers & Education*, 151, 103855. <https://doi.org/10.1016/j.compedu.2020.103855>
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020b). Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- Chen, X., Zou, D., Xie, H., & Cheng, G. (2021). Twenty years of personalized language learning: Topic modeling and knowledge mapping. *Educational Technology & Society*, 24(1), 205–222.
- Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two Decades of Artificial Intelligence in Education: Contributors, Collaborations, Research Topics, Challenges, and Future Directions. *Educational Technology & Society*, 25(1), 28–48.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An Approach for detecting, quantifying, and visualizing the evolution of a research field: A Practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, 5(1), 146–166. <https://doi.org/10.1016/j.joi.2010.10.002>
- Colpaert, J. (2006). Toward an ontological approach in goal-oriented language courseware design and its implications for technology-independent content structuring. *Computer Assisted Language Learning*, 19(2–3), 109–127. <https://doi.org/10.1080/09588220600821461>
- De Waal, F. (2009). *The Age of empathy: Nature’s lessons for a kinder society*. Three Rivers Press.

- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I, teacher: Using artificial intelligence (AI) and social robots in communication and instruction. *Communication Education*, 67(4), 473–480. <https://doi.org/10.1080/03634523.2018.1502459>
- Eguchi, A. (2017). Bringing robotics in classrooms. In *Robotics in STEM Education* (pp. 3–31). Springer. https://doi.org/10.1007/978-3-319-57786-9_1
- Fachada, N. (2018). Teaching database concepts to video game design and development students. *Revista Lusófona de Educação*, 40, 75–89. <https://core.ac.uk/download/pdf/270167972.pdf>
- Gosper, M., Green, D., McNeill, M., Phillips, R., Preston, G., & Woo, K. (2008). *The Impact of web-based lecture technologies on current and future practices in learning and teaching*. Australian Learning and Teaching Council. <http://www.cpd.mq.edu.au/teaching/wblt/research/report.html>
- Hart, S. A. (2016). Precision education initiative: Moving toward personalized education. *Mind, Brain, and Education*, 10(4), 209–211. <https://doi.org/10.1111/mbe.12109>
- Huang, Y. C., Cheng, Y. T., Chen, L. L., & Hsu, J. Y. J. (2019). Human-AI co-learning for data-driven AI. *PsyArXiv*. <https://doi.org/10.48550/arXiv.1910.12544>
- Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A Systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- Kerly, A., Ellis, R., & Bull, S. (2008). CALMsystem: A Conversational agent for learner modelling. *Knowledge-Based Systems*, 21(3), 238–246. https://doi.org/10.1007/978-1-84800-086-5_7
- Kim, N.-Y., Cha, Y., & Kim, H.-S. (2019). Future English learning: Chatbots and artificial intelligence. *Multimedia-Assisted Language Learning*, 22(3), 32–53. <https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=ec2a4549-8d90-4516-9864-0163f81dfe90%40redis>
- Kubilinskienė, S., Žilinskienė, I., Dagienė, V., & Sinkevičius, V. (2017). Applying robotics in school education: A Systematic review. *Baltic Journal of Modern Computing*, 5(1), 50–69. <https://doi.org/10.22364/bjmc.2017.5.1.04>
- Lee, M., Tran, D. T., & Lee, J.-H. (2021). 3D facial pain expression for a care training assistant robot in an elderly care education environment. *Frontiers in Robotics and AI*, 8, 42. <https://doi.org/10.3389/frobt.2021.632015>
- Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2021). Trustworthy AI: From principles to practices. *PsyArXiv*. <https://doi.org/10.48550/arXiv.2110.01167>
- López-Robles, J. R., Otegi-Olaso, J. R., Gómez, I. P., & Cobo, M. J. (2019). 30 years of intelligence models in management and business: A Bibliometric review. *International Journal of Information Management*, 48, 22–38. <https://doi.org/10.1016/j.ijinfomgt.2019.01.013>
- Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, 21(2), 220–232.
- Majid, R., & Santoso, H. A. (2021). Conversations sentiment and intent categorization using context RNN for emotion recognition. In *7th ICACCS*, 1, 46–50. <https://doi.org/10.1109/ICACCS51430.2021.9441740>
- Morton, H., & Jack, M. A. (2005). Scenario-based spoken interaction with virtual agents. *Computer Assisted Language Learning*, 18(3), 171–191. <https://doi.org/10.1080/09588220500173344>
- Moyi, P. (2019). Education for children with disabilities: will policy changes promote equal access in Kenya? *Comparative and International Education*, 47(2), 1–15. <https://doi.org/10.5206/cie-eci.v47i2.9329>
- Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., & Strappelli, L. (2022). Towards involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems*. <https://dl.acm.org/doi/pdf/10.1145/3514258>
- Oliver, M. (2011). Technological determinism in educational technology research: Some alternative ways of thinking about the relationship between learning and technology. *Journal of Computer Assisted Learning*, 27(5), 373–384. <https://doi.org/10.1111/j.1365-2729.2011.00406.x>
- Rajkumar, R., & Ganapathy, V. (2020). Bio-inspiring learning style chatbot inventory using brain computing interface to increase the efficiency of e-learning. *IEEE Access*, 8, 67377–67395. <https://doi.org/10.1109/ACCESS.2020.2984591>
- Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11(5), 5–16. <http://doi.org/10.22215/timreview/1438>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(1), 1–40. <http://doi.org/10.18637/jss.v091.i02>

- Santos, K.-A., Ong, E., & Resurreccion, R. (2020). Therapist vibe: Children's expressions of their emotions through storytelling with a chatbot. In *19th ACM IDC*, 483–494. <https://doi.org/10.1145/3392063.3394405>
- Seering, J., Luria, M., Ye, C., Kaufman, G., & Hammer, J. (2020). It takes a village: Integrating an adaptive chatbot into an online gaming community. *CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3313831.3376708>
- She, T., & Ren, F. (2021). Enhance the language ability of humanoid robot NAO through deep learning to interact with autistic children. *Electronics*, 10(19), 2393. <https://doi.org/10.3390/electronics10192393>
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31. <https://doi.org/10.1145/3419764>
- Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A Review of educational chatbots for the Facebook Messenger. *Computers & Education*, 151, 103862. <https://doi.org/10.1016/j.compedu.2020.103862>
- Sreelakshmi, A. S., Abhinaya, S. B., Nair, A., & Nirmala, S. J. (2019). A Question answering and quiz generation chatbot for education. In *2019 GHCI*, 1–6. <https://doi.org/10.1109/GHCI47972.2019.9071832>
- Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y., Dong, J., Duffy, V. G., Fang, X., Fidopiastis, G., Fragomeni, G., Fu, L. P., Guo, Y., Harris, D., Ioannou, A., Jeong, K., Konomi, S., Krömker, H., Kurosu, M., Lewis, J. R., Marcus, A., Meiselwitz, G., Moallem, A., Mori, H., Nah, F. F.-H., Ntoa, S., Rau, P.-L. P., Schmorow, D., Siau, K., Streitz, N., Wang, W., Yamamoto, S., Zaphiris, P., & Zhou, J. (2019). Seven HCI grand challenges. *International Journal of Human-Computer Interaction*, 35(14), 1229–1269. <https://doi.org/10.1080/10447318.2019.1619259>
- Tegos, S., Demetriadis, S., & Karakostas, A. (2011). MentorChat: Introducing a configurable conversational agent as a tool for adaptive online collaboration support. In *15th Panhellenic Conference on Informatics* (pp. 13–17). IEEE. <https://doi.org/10.1109/PCI.2011.24>
- Tlili, A., Lin, V., Chen, N.-S., & Huang, R. (2020). A Systematic review on robot-assisted special education from the activity theory perspective. *Educational Technology & Society*, 23(3), 95–109.
- Vijayan, A., Janmasree, S., Keerthana, C., & Sylva, L. B. (2018). A Framework for intelligent learning assistant platform based on cognitive computing for children with autism spectrum disorder. In *International CET Conference on IC4* (pp. 361–365). <https://doi.org/10.1109/CETIC4.2018.8530940>
- Vygotsky, L. S. (1978). *Mind in society: The Development of higher psychological processes*. Harvard University Press.
- Wambsganss, T., & Rietsche, R. (2019). Towards designing an adaptive argumentation learning tool. In *Proceedings of the International Conference on Information Systems (ICIS) 2019*. <https://www.alexandria.unisg.ch/publications/259195>
- Weitekamp, D., Harpstead, E., & Koedinger, K. R. (2020). An Interaction design for machine teaching to develop AI tutors. *CHI Conference on Human Factors in Computing Systems*, 1–11. <https://doi.org/10.1145/3313831.3376226>
- Williams, R., Park, H. W., Oh, L., & Breazeal, C. (2019). Popbots: Designing an artificial intelligence curriculum for early childhood education. *AAAI Conference on Artificial Intelligence*, 33(01), 9729–9736. <https://doi.org/10.1609/aaai.v33i01.33019729>
- Xia, L., & Zhong, B. (2018). A Systematic review on teaching and learning robotics content knowledge in K-12. *Computers & Education*, 127, 267–282. <https://doi.org/10.1016/j.compedu.2018.09.007>
- Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2022). Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human-Computer Interaction*, 1–25. <https://doi.org/10.1080/10447318.2022.2041900>
- Yang, S. J. H. (2021). Guest editorial: Precision education-a new challenge for AI in education. *Educational Technology & Society*, 24(1), 105–108.
- Yang, S. J. H., Ogata, H., & Matsui, T. (2023). Guest editorial: Human-centered AI in education: Augment human intelligence with machine intelligence. *Educational Technology & Society*, 26(1), 95-98.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.-S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1–27. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhong, B., Zheng, J., & Zhan, Z. (2020). An Exploration of combining virtual and physical robots in robotics education. *Interactive Learning Environments*, 1–13. <https://doi.org/10.1080/10494820.2020.1786409>

A Risk Framework for Human-centered Artificial Intelligence in Education: Based on Literature Review and Delphi–AHP Method

Shijin Li and Xiaoqing Gu*

Department of Education Information Technology, East China Normal University, China // shijinliEdu@163.com // xqgu@ses.ecnu.edu.cn

*Corresponding author

ABSTRACT: With artificial intelligence (AI) is extensively applied in education, human-centered AI (HCAI) has become an active field. There although has been increasing concern about how to systematically enhance the AI applications effect, AI risk governance in HCAI education has not been discussed yet. This study adopted literature meta-analysis, along with the Delphi and analytic hierarchy process (AHP) methods in order to establish the risk framework and calculate the index weight of HCAI education. The results confirm that the risk framework includes eight indicators, which respectively are misunderstanding of the HCAI concept (MC), misuse of AI resources (MR), mismatching of AI pedagogy (MP), privacy security risk (PSR), transparency risk (TR), accountability risk (AR), bias risk (BR), and perceived risk (PR). Meanwhile, the eight indicators are divided into four categories such as HCAI concept, application process, ethical security, and man-machine interaction. Moreover, the trend of risks types indicates that more than half of the articles consider only three or less risks types, and the evolution results of risks indicators gradually increased between 2010 and 2021. Additionally, the weights of the eight indicators are $MP > MR > AR > PSR > TR > PR > BR > MC$. Results obtained could provide theoretical evidence and development suggestions for future scientific governance of HCAI education. Furthermore, the risk framework not only systematically considers the risk governance order of HCAI education, but more importantly, it is the key bridge to the collaborative advancement of stakeholders such as managers, teachers, students, and parents, which can contribute to the scientific, healthy, and sustainable HCAI education.

Keywords: Human-centered artificial intelligence (HCAI), Risk framework, Index weight, AHP, Delphi

1. Introduction

With data analysis and autonomous learning, artificial intelligence in education (AIED) applications have been making a wider impact on personalized learning, classroom monitoring, student performance, sentiment analysis and decision evaluation (Hwang et al., 2020). For example, intelligent tutors and virtual learning partners can help students perform communication and cooperative tasks independently and efficiently (Holmes et al., 2019). Adaptive learning systems can provide adaptive feedback and service support (Chin & Tseng, 2021). Automatic question-answer technology can solve students' classroom problems in real time (Lu et al., 2021; Perikos et al., 2017). Emotion detection technology can dynamically perceive students' emotional needs and provide personalized emotional support (Chen et al., 2021; Saneiro et al., 2014). Decision management technology can automatically diagnose students' learning needs and assist them in decision-making (Yang et al., 2021). In summary, AI has become a key point in empowering and transforming education, and AIED applications will be developed at a large scale (Zawacki-Richter et al., 2019; UNESCO, 2019). Similar to the “dual-use” nature of biochemical technologies, AIED applications offer both rewards and potential risks. With proper use of AI, it can improve the human condition for education in many ways, but the misuse of AI due to a range of risks (White & Lidskog, 2022). Therefore, the risk governance framework must be developed to ensure the responsible and sustainable of AIED applications.

Human-centered AI (HCAI) is one effective approach that holds promise for the responsible AIED applications, as well as systematically consider AI algorithms through humanistic situation, thereby enhancing human intelligence rather than replace them with machines. Stanford University, UC Berkeley and MIT have set up HCAI research institutes, aiming to develop humanistic, ethical, and beneficial AI education. Researchers have begun to discuss ethical design approaches, but AI risk governance in HCAI education has not been discussed yet. More importantly, the risks of HCAI education are highly complex, unpredictable, and nonlinear (Renn, 2021), and without an overall framework, it is difficult to systematically identify, understand and manage risks (Schweizer, 2021). Although previous studies have reviewed algorithm bias (Kusner & Loftus, 2020), technology abuse (Jim & Chang, 2018), privacy security (Sivill, 2019), and role ambiguity (Guilherme, 2019), but there is no systematic risk framework for HCAI education. Therefore, it is necessary to put forward the risk framework as well as index weight of HCAI education. In order to advocate the idea of HCAI, implement the method of AI under human-control and avoid potential negative effects, the study adopted literature meta-

analysis, along with the Delphi and analytic hierarchy process (AHP) methods, and established the risk framework and calculate the index weight. The main objective of this study is to solve how to systematically govern risks and help stakeholders obtain optimal benefits while adopting forward-looking actions. In addition, we can implement responsible, sustainable, and healthy HCAI education based on the risk framework. In particular, this study offers a reference risk regulatory framework of HCAI education, which can contribute to enhancing the practice effects and application benefits.

2. Literature review

2.1. Responsible AIED: HCAI research and discovery

HCAI is an ideological paradigm that places humans at the center of the man-machine collaboration paradigm, abiding by the ethics, common values, and interests of human beings. Different research teams have also carried out a series of discussions, aiming to introduce the HCAI concept into the design and practice process, so as to promote the sustainable and responsible AIED applications. Shneiderman (2020b) visually described HCAI as the “AI Copernican Revolution,” and profoundly expounded on the HCAI concept and widely advocated the use of humanistic algorithms for design, development and application. Schmidt (2020) argued that HCAI was designed with a clear purpose for human benefit, while being transparent about who had control over data and algorithms. Xu (2019) proposed an extended HCAI framework that included ethically aligned design, technology enhancement and human factor design, so as to ensure AI solutions are explainable and comprehensible.

HCAI emphasizes the integration of human role into the human-machine system, and develops human-machine hybrid enhanced intelligence through the complementation of human-machine intelligence. Nowadays, the research progresses of HCAI domain mainly focus on human intelligence enhancement, human-machine hybrid enhanced intelligence, human-AI cooperation, explicable AI, human-controllable autonomy, intelligent human-machine interaction, and ethical AI design (Xu et al., 2021). In particular, ethical AI design is an important issue in HCAI education, and it is also the basis for achieving the HCAI goals. Moreover, without an ethical AI design framework, the HCAI concept cannot be realized, and safe, reliable, and trustworthy AI systems cannot be developed. Therefore, an important task of HCAI research is to develop AI risk governance framework.

2.2. AIED risk governance as a scientific way to realize HCAI education

In 2015, google image software labeled a black African-American couple as “gorilla,” which not only showed the poor performance of the model in face recognition, but more importantly the lack of basic respect for colored race (Benjamin, 2019). A Princeton university study emphasize that the biased AI algorithm link women with “family” and “art,” men with “career” and “ambition,” and link colored race with unpleasant words (Caliskan et al., 2017). Angwin et al. (2016) and Kay et al. (2015) exposed gender and racial biases in career development and predictive education systems. According to Ahn et al. (2021), intelligent agents can automatically obtain students’ learning styles, habits, and abilities. However, if the AI systems predict that students will fail in the next exam according to student behavior data during a certain period, will it affect students’ self-confidence? In addition, questions such as who can own data or whether the data are real and valid are common risks in AIED applications (Ketamo, 2018).

Whenever a new technology appears, we are always eager to put it into use for fear of missing its educational benefits, which often leads to a series of risks. The AIED risk governance has become a social consensus (Floridi et al., 2018). Different research teams also conducted a series of systematic reviews and pointed out practical problems (Deeva et al., 2021; Winters et al., 2020; Scherer, 2015; Jim & Chang, 2018), which are (1) How to define and dispose of the new roles of teachers and their relationship with intelligent systems? (2) How can students’ privacy safety be protected when collecting behavioral data? (3) What kind of ethical knowledge should stakeholders possess and what ethical criteria should they follow? (4) Whose interests should AI give priority to conflicting stakeholders? (5) If AI learning fails, who should be held accountable? In fact, an accountability system accompanies the entire life cycle of the AI systems, and responsible AI systems can be constructed through access regulations, timely supervision, and decision-making evaluation. In this process, it is difficult to identify the party responsible. This might because the responsibility is based on free will, but machines do not have free will, in this way, social structural barriers and personal cognitive barriers lead to design, data, and algorithm biases in AIED applications. Furthermore, the responsibility for defects in intelligent products cannot be completely transferred to manufacturers, nor can designers and programmers be absolved in the AI systems development process.

AIED risk governance is the scientific way to realize HCAI education. However, there are three deficiencies at present, first, the research perspectives are mostly theoretical exploration of risks characterization, and there is no research to systematically consider the risk framework in HCAI education. Moreover, research method is mainly the literature or the survey method, and there is no method for calculating the risk weight. Additionally, the guidance, operability and extensibility of research conclusions need to be improved urgently.

2.3. Purpose of the current study

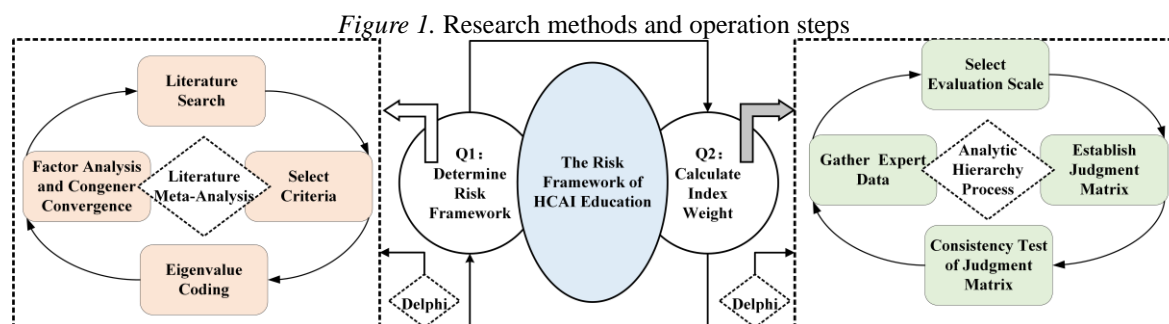
Since AIED risks are not only AI technical issues, but also involve the relationship between education and society, so it is necessary to integrate the characteristics of “Technology-Education-Society,” and systematically consider the risk framework and index weight. Delphi-AHP is a qualitative and quantitative decision-making method (Turón et al., 2019), by collecting, summarizing and analyzing the relative importance of experts to each index, using the AHP method to determine the index weight, and combining qualitative and quantitative evidence feedback, an operable theory framework and index weight are finally formed. Based on this, we used Delphi-AHP to develop the risk framework and index weight of HCAI education. Through the analysis of index meaning and weight level, which not only provide theoretical evidence for risks governance, but also enhance practical guidance for the design of risks intervention programs and the development of risks assessment tools.

This study aims to answer the following four problems:

- What indicators are included in the risk framework of HCAI education? And what are characteristics of each risk?
- What is the trend of risks types in HCAI education?
- What are the evolution results of risks indicators in HCAI education?
- What are the weights of these indicators?

3. Methodology

The purpose of this study is to develop the risk framework and establish index weight of HCAI education. To achieve the aims, the following research methods and operation steps are designed based on systematic principles (see Figure 1).



3.1. Literature meta-analysis method

To solve the first research question, the literature meta-analysis method was used to determine the risk framework, which followed the process of “literature search → select criteria → eigenvalue coding → factor analysis and congener convergence.” The literature search terms were conducted with reference to previous studies (e.g., “HCAI” in Shneiderman (2020a) and “ethical framework” in Floridi et al. (2018)) by considering both HCAI and risk fields.

The processes of literature meta-analysis are as follows:

- The academic databases used to collect articles are Web of Science, Scopus, Science Direct, EBSCO, Wiley Online Library, ProQuest, ACM, IEEE and Google Scholar.
- The keywords used for literature search are (“artificial intelligence” OR “AI” OR “Human-centered artificial intelligence” OR “HCAI” OR “AIED” OR “AIED”) AND (“risk” OR “risk framework”).
- The time range of articles published from January 2010 to December 2021, as AIED applications have become widely popular since 2010.

- The selected articles are used to develop the risk framework of HCAI education, and the selection criteria mainly consider the following two points, one is the research context is AI education, another is the research topic includes risk types. When one of the selection criteria was not met, the article was excluded. According to the above selection criteria, 50 valid samples were finally obtained.
- In the process of eigenvalue coding, we focus on what types of risks are included in the literature? And what are the significant or potential features of risks? Through factor analysis and congener convergence, and after two rounds of expert consultation, we finally developed the risk framework of HCAI education.

3.2. Delphi-AHP method

To solve the second research question, the Delphi and AHP methods were used to calculate index weight. Expert groups directly determine the content of consultation and the validity of data results (Goodman, 1987). In order to ensure the scientificity and validity of the research samples, the study adopted a combination of cluster sampling and convenience sampling to determine the expert groups (Etikan & Bala, 2017; Cohen et al., 2017). First, we used the cluster sampling method, and took 147 double-first-class universities in China as the first-level sampling units. Then, the convenience sampling method was used to select expert groups that could meet the research needs. Additionally, three selection criteria were set throughout the sampling process: (1) Very familiar or relatively familiar with the research topics of “HCAI education” and “AIED risk.” (2) The work unit is a double first-class university. (3) Both domestic and foreign multicultural background. Based on this, we finally identified 37 experts who traversed 10 universities in eastern, central and western China (see Table 1).

Table 1. Basic information statistics of 37 experts

Basic information of experts		Number	Proportion
Gender	Male	29	78.4%
	Female	8	21.6%
Multicultural background	Study abroad experience	25	67.6%
	International exchange program	12	32.4%
	Tsinghua University	2	5.4%
	Beijing University	2	5.4%
	Beijing Normal University	4	10.8%
	East China Normal University	6	16.2%
	Zhejiang University	2	5.4%
Work units	Central China Normal University	5	13.6%
	Shaanxi Normal University	6	16.2%
	Southwest University	3	8.1%
	South China Normal University	3	8.1%
	Nanjing Normal University	4	10.8%

Meanwhile, in order to ensure the objectivity of data samples, the level judgment of expert authority (Cr) is added, that is from the judgment basis (Ca) and familiarity (Cs) comprehensively consider the data results of experts (see Table 2). According to the calculation formula, $Cr = (Ca + Cs)/2$, 37 experts' judgment basis (Ca) is $(35*0.5+2*0.4+32*0.3+5*0.2+34*0.1+3*0.1+25*0.1+5*0.1+7*0.1)/37 = 0.98$. The degree of familiarity (Cs) is $(30*1.0+7*0.8)/37 = 0.96$. Thus, expert authority (Cr) is $(0.98 + 0.96)/2 = 0.97$. Since the degree of expert authority $(Cr) \geq 0.7$, the results of expert consultation are reliable.

Table 2. Expert authority and weight coefficient

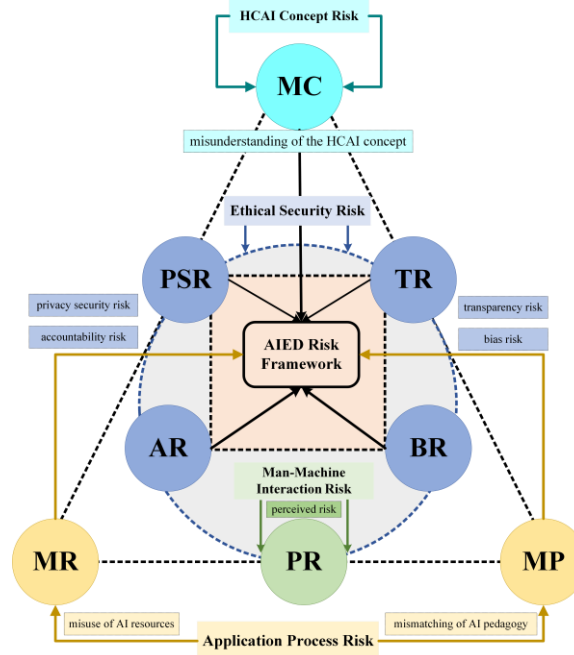
Judgment basis and weight coefficient					
Judgment basis	Large		Medium		Small
Practical experience	0.5		0.4		0.3
Theoretical analysis	0.3		0.2		0.1
Peer understanding	0.1		0.1		0.1
Intuitive feeling	0.1		0.1		0.1
Familiarity and weight coefficient					
Familiarity	Very familiar	Familiar	General Familiar	Not very familiar	Unfamiliar
Weight coefficient	1.0	0.8	0.6	0.4	0.2

4. Results

4.1. Analysis of the risk framework structure of HCAI education

Through literature meta-analysis and two rounds of Delphi, we finally determined the risk framework of HCAI education (See Figure 2), which includes misunderstanding of the HCAI concept (MC), misuse of AI resources (MR), mismatching of AI pedagogy (MP), privacy security risk (PSR), transparency risk (TR), accountability risk (AR), bias risk (BR), and perceived risk (PR). Meanwhile, these eight indicators are divided into four categories such as HCAI concept, application process, ethical security, and man-machine interaction.

Figure 2. The risk framework for HCAI in education



Based on our results, HCAI concept risk includes MC, application process risks include MR and MP, ethical security risks include PSR, TR, AR and BR, and man-machine interaction risk includes PR. In particular, intelligent concept risk stems from the ontological risk of ignoring AI technology to restore education world, application process risk originates from the cognitive risk of masking AI technology to characterize education ecology, ethical security risk stems from the value risk that neglecting AI technology goes against the original intention of education, man-machine interaction risk stems from the ethical risk of education governance caused by the misuse of AI technology.

4.2. Analysis of the trend of risks types in HCAI education

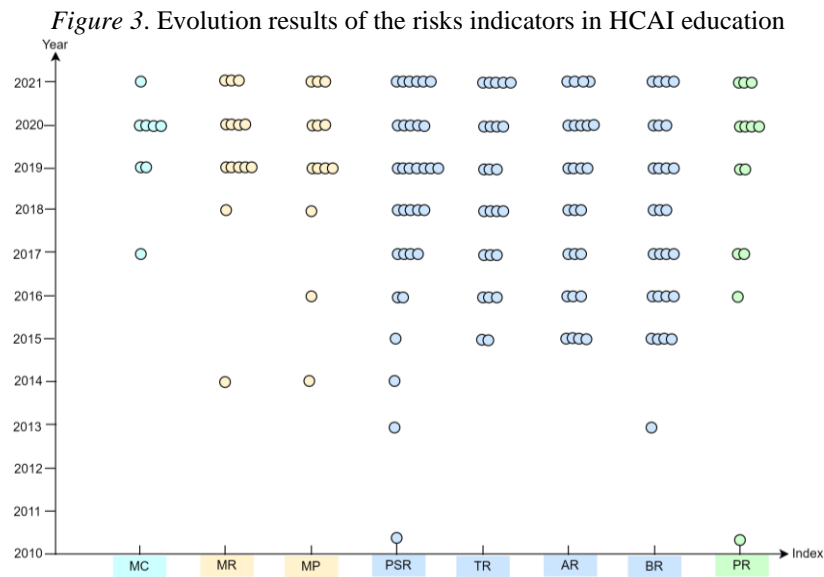
Table 3 shows the trend of risks types. The eight indicators are distributed in 50 articles. The top one risk index accounted for 64% of the total articles. The top three articles ranked by number of indicators are included seven indicators, the first article focuses on the risks in AIED applications process, and the last two specialize in the risks and challenges of AIED. Among the articles listed, BR (28), AR (26) and TR (24) are almost equally numerous. Meanwhile, more than half of the articles consider only three or less risks types.

Table 3. The trend of risks types in HCAI education

Indicators	Citation	Brief description of the research	MC	MR	MP	PSR	TR	AR	BR	PR
			8	14	13	32	24	26	28	13
7	Zhang, 2021	The reform and innovation of AI technology for information service	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7	Hwang et al., 2020	Vision, challenges, roles and research issues of AIED	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7	UNESCO, 2019	Challenges and opportunities for sustainable	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

		development of AIED								
6	Renz & Vladova, 2021	HCAI in educational technologies	Yes			Yes	Yes	Yes	Yes	Yes
5	Xu, 2019	HCAI from interaction aspect	Yes	Yes	Yes	Yes				Yes
5	Floridi et al., 2018	An ethical framework (AI4People) for a good AI society		Yes		Yes	Yes	Yes	Yes	
5	Caliskan et al., 2017	Bias in humans and machines				Yes	Yes	Yes	Yes	Yes
4	White & Lidskog, 2022	Ignorance and the regulation of AI technology				Yes	Yes	Yes	Yes	
4	Ahn et al., 2021	Privacy, transparency and trust in K-12 learning analytics				Yes	Yes	Yes	Yes	
4	Deeva et al., 2021	Automated feedback systems for learners		Yes	Yes	Yes	Yes			
4	Wu et al., 2020	Ethical principles and governance process of AI technology				Yes	Yes	Yes	Yes	
4	Sivill, 2019	Ethical and statistical considerations in models of moral judgments				Yes	Yes	Yes	Yes	
4	Intel Corporation, 2018	Individuals' privacy and data in the AI world				Yes	Yes	Yes	Yes	
4	Jim & Chang, 2018	Data governance in higher education				Yes	Yes	Yes	Yes	
4	Boddington, 2017	Ethics for artificial intelligence				Yes	Yes	Yes	Yes	
4	Wessels, 2015	Authentication, status, and power in a digitally organized society				Yes	Yes	Yes	Yes	
3	Winters et al., 2020	Digital structural violence in future learning systems					Yes	Yes	Yes	
3	Chen et al., 2020	Application and theory gaps in AIED	Yes	Yes	Yes					
3	Auernhammer, 2020	HCAI design framework	Yes			Yes				Yes
3	Kusner & Loftus, 2020	Conceptual paper on the fairer algorithms					Yes	Yes	Yes	
3	Zawacki-Richter et al., 2019	AI applications in higher education		Yes	Yes	Yes				
3	Friedman et al., 2017	A survey of value sensitive design methods	Yes			Yes				Yes
3	Kitchin, 2017	Thinking critically about and researching algorithms					Yes	Yes	Yes	
3	OECD, 2016	The impact of digital technologies on teaching and learning			Yes	Yes				Yes
3	Mittelstadt et al., 2016	The ethics of algorithms					Yes	Yes	Yes	
3	Ozga, 2016	Digital data use in education					Yes	Yes	Yes	
3	Burrell, 2016	The opacity in machine learning algorithms					Yes	Yes	Yes	
3	Pasquale, 2015	The black box society					Yes	Yes	Yes	
3	Chang et al., 2014	Augmented reality versus interactive simulation technology		Yes	Yes	Yes				

risks indicators were appeared simultaneously. Additionally, ethical security risks like PSR, TR, AR, BR are always the focus of AIED applications.



4.4. Results of index weight

According to the analysis process of “establish judgment matrix → consistency test of judgment matrix → gather expert data,” we used the AHP method and Yaahp software to calculate the weights of eight risks indicators.

The first is the establish judgment matrix, in this process, the key is to select evaluation scale. AHP method usually uses the nine-level evaluation to judge index factors in pairs (Saaty, 1987). This is because the limit of the difference between the two objects is 7 ± 2 . Therefore, in order to eliminate errors as much as possible, we selected the classic nine-level evaluation method to compare the importance of indicators in pairs (see Table 4). In the specific operation process, 37 experts used a nine-level evaluation method to judge the relative importance of eight indicators, and eight judgment matrices were established in Yaahp software for eight risks indicators.

Table 4. Evaluation method of judgment matrix

Scale	Definition	Connotation
1	Equally important	The two elements are of equal importance
3	Slightly important	Compared with the two elements, the former is slightly more important than the latter
5	Quite important	Compared with the two elements, the former is quite important than the latter
7	Obviously important	Compared with the two elements, the former is obviously more important than the latter
9	Absolutely important	Compared with the two elements, the former is absolutely more important than the latter
2,4,6,8	—	Indicates an intermediate value between the above criteria
Reciprocal of 1~9	—	Indicates the importance of the comparison of the corresponding two-factor exchange order

The second is the consistency test of judgment matrix. In this process, the minimum algorithm was used for automatic correction. After correction, the judgment matrices of 37 experts all met the statistical standard of consistency ratio $CR < 0.1$. Then, 37 judgment matrices corresponding to each expert's information were formed in Yaahp software, based on which an aggregated judgment matrix was formed (see Table 5).

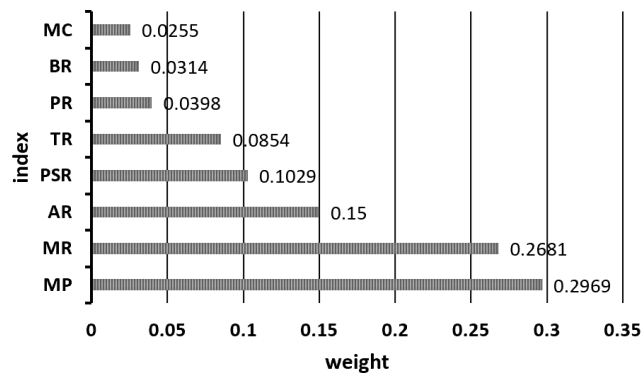
The third is the gather expert data, which includes two methods of calculation result aggregation and judgment matrix aggregation. The former calculates the average of the ranking weights obtained by each expert judgment matrix as the aggregation result, and the latter takes the average of the expert judgment matrix results and calculates the ranking index weight. Even if 37 experts' judgment matrices meet the consistency requirements, the final results obtained after the combined judgment matrices are also likely to have some problems, like the

individual and the group judgment matrix have inconsistent meanings, and lack of data semantics. Therefore, we used the calculation result aggregation to output weights of eight risks indicators (see Figure 4).

Table 5. Aggregated judgment matrix of 37 experts

AIED risk	MC	MR	MP	PSR	TK	AR	BR	PR
MC	1	0.0953	0.0861	0.2482	0.2992	0.1704	0.8142	0.6422
MR	10.4946	1	0.9032	2.6053	3.1404	1.7878	8.5452	6.7396
MP	11.6198	1.1072	1	2.8846	3.4771	1.9794	9.4614	7.4623
PSR	4.0282	0.3838	0.3467	1	1.2054	0.6862	3.2800	2.5869
TK	3.3418	0.3184	0.2876	0.8296	1	0.5693	2.7211	2.1461
AR	5.8703	0.5594	0.5052	1.4573	1.7566	1	4.7799	3.7699
BR	1.2281	0.1170	0.1057	0.3049	0.3675	0.2092	1	0.7887
PR	1.5571	0.1484	0.1340	0.3866	0.4660	0.2653	1.2679	1

Figure 4. Weights of eight risks indicators



5. Discussion

In this study, we used the literature meta-analysis method to systematically develop eight risk indicators in HCAI education, which were divided into four categories of risks such as HCAI concept (MC), application process (MR, MP), ethical security (PSR, TR, AR, BR) and man-machine interaction (PR). Meanwhile, we used the Delphi and AHP methods to calculate the weights of eight indicators, which were $MP > MR > AR > PSR > TR > PR > BR > MC$. Furthermore, such a framework provides theoretical reference standard for the risk governance in HCAI education. Findings regarding risk framework and weights analyses provide profound insight for future HCAI education research, as described in the following subsections from large to small weights.

5.1. The MP and weight analysis

Innovative pedagogy is the key of the AIED application process. That would mean if AI pedagogy is not innovated in time and not adequately prepared for the potential of AI technology, the AIED practice effect may be more harmful than beneficial. Harri Ketamo, an AI researcher who held the same view, pointed out that “learning is hard work, but we can make learning more enjoyable, easier and effective through good pedagogy” (Ketamo, 2018). Moreover, Sharples (2019) argued that the key to innovative teaching lies in how to construct a pedagogy-technology fit. To solve this, Lu et al. (2021) proposed that the school management level should form AI interschool alliances and explore pedagogy-technology fit through expert support or case studies. Furthermore, according to Chen et al. (2021), we can use innovative pedagogies such as chat robots and remote collaborative learning to strengthen learners’ knowledge about constructive, social, and contextual understanding, also promote the continuous excitation of inquiry motivation and intelligent emotion.

Our research found that MP is the biggest risk in HCAI education. Therefore, in order to prevent AIED applications from falling into the dilemma of “wearing new shoes and walking the old road,” it is important to pay attention to pedagogy-technology fit. However, as AI courses are mostly used as an elective or school-based curriculum, the curriculum coherence of each semester is also insufficient. In addition, the teaching materials, teaching concepts, and intelligent tools of different schools are quite different, which generally leads to unsystematic AIED pedagogy design (Zhang et al., 2021). Thus, future research may consider exploring HCAI teaching practice based on innovative pedagogy-AI technology fit.

5.2. The MR and weight analysis

School-based resources are the foundation of the AIED application process, and if the intelligent resources are unreliable or invalid, which will lead to poor AI learning effects. Our results showed that MR is the second risk in HCAI education. This might include the following three reasons: First, AI resources are complex and cluttered, because AI resources are not specifically targeted at education activities, so they are not directly meet the AIED applications. Second, AI resources generally lack systematic course design and resource construction, which also lead to the incoherence of intelligent resources between each learning section. Third, the contents of AI resources are differentiated, and a large number of AI resources in the exploratory stage or esoteric have entered the classroom. If schools fail to transform intelligent resources in time, the AIED practice will fall into the misunderstanding of blindly “seeking innovation” or “seeking perfection.” Therefore, K-12 schools need to tailor, adjust, arrange, and even re-develop existing AI resources, so that AI resources can adapt to different teaching scenarios.

According to Holmes et al. (2019), from the perspective of intelligent resource design, the primary task is to build school-based intelligent resources, and use AI technologies such as big data, deep learning and knowledge graph to open up AI resource-sharing platforms in different regions and schools. From the perspective of intelligent resource linkage, we can establish collaborative mechanisms and scientifically adjust the allocation plan of AI resources based on regional structure, investment level, dynamic mechanism, power, and responsibility. Meanwhile, the whole process should also provide corresponding regulatory measures and institutional guarantees. Furthermore, we recommend that future research should focus on the three forms of risks: The first are source risks, such as the convergence, sharing, and circulation mechanisms of social AI resources; the second are process risks, such as identifiable, traceable, decentralized, and transparent; and the third are port risks, such as certification standards and evaluation indicators of school-based AI social resources.

5.3. The AR and weight analysis

AR is one of ethical security risks in AIED applications. In Boddington (2017), since intelligent algorithms failed to understand the real cause of risks, and this also led to the ambiguity of responsibility. Orr and Davis (2020) also emphasized that AR determination with AI technology is not easy. Specifically, AI systems do not have the ability to bear legal responsibility independently, so the accountability mechanism is meaningless to some degree. Moreover, AI products have the ability of autonomous and independent learning, judgment, and decision-making, product designers and program developers cannot fully govern the evolutionary behavior of AI products, so it is difficult to plan the possible adverse consequences in advance. In addition, there is a lack of effective accountable design methodologies or technical details to guide design specification about the AR in AIED applications.

Our research found that AR is the third risk in HCAI education. However, little considerations are given to how to effectively clarify responsibilities and normative criterion, and most developers even consider responsibility design at a later stage rather than during the development process of AI systems. Therefore, future research may focus on establishing accountability mechanism from the key links such as technical design and institutional guarantee. Also, we recommend that more studies consider implementing accountability in AI systems design based on HCAI approach, and taking systematic and effective measures in design, testing and professional training.

5.4. The PSR and weight analysis

PSR is that may expose personal privacy and personalized needs in AI applications, which belongs to one of the ethical security risks. Zhang (2021) argued that the breach of data privacy is eroding the well-being of learners. For example, the information leakage caused by the head ring, the labeling of learning evidence or the hybridization of heterogeneous data. In this way, AIED applications are releasing a lot of privacy security through procedures and rules, as the Foucault-style “panoramic prison.” Moreover, when learners use AI technology for a long time, it is easy to develop the bad habit of “technical flow.” That is to say, once learners are out of the technical cage, they will avoid the cooperation and communication between peers, and then produce undesirable symptoms such as withdrawn temperament and emotional alienation.

Based on the results, PSR is the fourth risk in HCAI education. Nevertheless, privacy security runs through the whole process of AIED applications. Thus, more studies may comprehensively consider the PSR combined with different scenarios. For example, at the individual level, AI systems must fully focus on the privacy protection of

independent personal data, mobile data, names and so on (Zhang et al., 2021). At the collective level, since AI systems are likely to collect and utilize group information illegally by stealing, tampering, and leaking, so future research should focus on data flow and interaction specification. In addition, it is also necessary to establish blockchain trust mechanisms and data regulatory agencies to supervise data collection, legal use and privacy security.

5.5. The TR and weight analysis

TR represents AI technology cannot provide sufficient explanatory information. In Mittelstadt et al. (2016), most explainable AI projects are carried out only within the AI discipline. Also, some AI personnel adopted an “algorithm-centric” approach, and even built explainable AI for themselves rather than users, which exacerbated the opacity of algorithms. In this way, AI technology process and implementation details are often hidden, the packaging characteristics of the “black box” create a near “perfect illusion” for AIED applications, making it difficult for stakeholders to grasp the actual differences between data and entities (Burrell, 2016; Kitchin, 2017; Ozga, 2016). Subsequently, explainable AI (XAI) has become a research hotspot. For example, develop or improve ML technology to obtain interpretable algorithmic models. Also, develop XAI of user models with the help of advanced human-machine interaction technology. Furthermore, evaluate specific psychological explanation theories to assist in the development of XAI.

In our research, TR is the fifth risk in HCAI education. This might because the lack of transparent design of AI systems, which affects the credibility of AIED applications. Thus, we should not only regard AI technology as an education tool, focusing on specific categories such as “why to teach” “who to teach” “what to teach” and “how to teach,” but should “apply to... no longer used for... ,” breaking through the shackles of “technology black box.” Moreover, if education information is transmitted in an understandable way, which can also enhance the fluidity, interactivity, and openness of XAI. Thus, future research may focus on developing XAI solutions based on HCAI concepts to meet AIED need.

5.6. The PR and weight analysis

PR is a combination of behavioral and environmental insecurity. In the era of AI, benefit trust and risk perception are interactive. In other words, the public’s subjective perception at the cognitive level can easily lead to panic or concern about privacy infringement. Specifically, Chatterjee and Bhattacharjee (2020) argued that when risk perception is high, individuals are less willing to adopt AI technology. Moreover, several findings also revealed that the lower the human-machine interaction risk, the more willing schools are to carry out AIED (Wang et al., 2021; Chai et al., 2020). In addition, if the effective communication in man-machine interaction can be enhanced, people’s perceived risks can be reduced in AIED applications. According to Xu (2019), man-machine interaction should pay attention to artificial stupidity (AS), because even a perfect computer program is nothing but a cold mechanism. This also shows that AS can stimulate the enthusiasm of human participation to some degree.

The man-machine interaction risk is ranked sixth. This might because when AI technology is integrated into education ecology, the multiple stakeholders of “home-school-society-enterprise” are prone to worry and panic that “intelligent tutors will replace human teachers” due to their lack of technical experience. In this situation, we should obtain systematic experience through literature meta-analysis to provide a basis for the human-machine interaction practice. Based on our results, there is still a lack of innovative human-machine collaborative teaching models, thus, we should set boundaries for man-machine interaction based on AS, and fully explore innovative models of balanced cooperation between machine intelligence and human intelligence. Meanwhile, future more studies may focus on reasonable and appropriate human-machine collaborative teaching process and evaluation technology, so as to build a new human-machine interaction ecology.

5.7. The BR and weight analysis

BR is the unfair attitude and biased judgment of a certain social group in advance. Knox et al. (2019) found that AI products intentionally excluded specific groups from the target audience, making it difficult for some learners to obtain equivalent education services. According to Nathanson et al. (2013), AI recommendation system did not achieve the goals of debiasing, which resulted in most of the low-achieving students being recommended to poor high schools. In particular, the algorithm is actually a “human concept embedded in mathematics,” the process follows the rule of “prejudice goes in, then prejudice goes out.” In this way, “filter bubble” can mislead

teachers' decisions, narrow students' minds and ideologies, and cause "echo chamber bias," "Matthew effect," "halo effect" and even "digital structural violence" in education (Wu et al., 2020).

Our research found that although BR is ranked second to last, we still need to widely expand educators' action awareness about BR in HCAI education. We thus suggest that stakeholders such as managers, researchers, and educational practitioners need to adopt collaborative innovation approach to maintain the dynamic balance and positive interaction in AIED applications. Meanwhile, scholars should keep up with the latest trends and components framework of PR, and comprehensively explore its dynamic mechanisms and avoidance strategies. In addition, future studies may consider exploring the PR models based on HCAI concept, so as to develop AI systems that are useful, usable, and in which humans have final control.

5.8. The MC and weight analysis

HCAI concept is the goal foundation of AIED applications. If the HCAI concept is widely integrated into AIED applications, moral values will become part of the AI systems design, so as to ensure the healthy, controllable and reliable AIED ecology. This might because the HCAI concept advocates the development of responsible AI education, which is crucial for establishing "high-quality and warm" AIED ecology. Also, this is consistent with the concept of human-in-the-loop (Honeycutt et al., 2020). According to Yang et al. (2021), human beings have features that are incomparable to AI in terms of cognition, emotion, attitude, and values. Verkijika et al. (2015) argued that it is possible to further explore enabling conditions for innovative learning and create effective intervention scaffolds from the perspective of human beings. For example, Dignum (2019) proposed that human value design and value-sensitive design (VSD), which put human rights, dignity, and freedom at the center of AI systems design, could identify, consider, and determine the adaptive path of man-machine collaboration.

Based on the results, although HCAI risk accounts for the smallest proportion in the risk framework, it is the primary index. Since it is consistent with the essential pursuit of HCAI education, which can also guarantee the integration of goals, processes, and results. In particular, the three forms of HCAI governance structure of reliable design, safe management, and credible certification can enhance public trust and confidence in AIED applications. Overall, the fundamental way to break through the AIED risk is to adhere to the HCAI concept and its endogenous laws. We thus suggest that more studies may consider designing, developing and applying HCAI-oriented practice paradigm.

6. Conclusions

Our study was the first-in-depth to explore the risk framework and establish index weight of HCAI education. To achieve the first aim, we used the literature meta-analysis method to determine the risk framework, and to achieve the second aim, we used the Delphi and AHP methods to calculate index weight. In sum, our study indicates that (1) the risk framework includes eight indicators, which are MC, MR, MP, PSR, TR, AR, BR, and PR; (2) eight indicators are divided into four categories such as HCAI concept, application process, ethical security, and man-machine interaction; (3) the trend of risks types confirms that more than half of the articles consider only three or less risks types; (4) the evolution results show that very limited risks indicators (e.g., PSR, TR, AR, BR) are considered before 2015, however, with the widespread increase of AIED applications, both the quantities and types of risks indicators (e.g., MC, MR, MP, PR) have increased in the last five years; (5) the weights of the eight indicators are $MP > MR > AR > PSR > TR > PR > BR > MC$.

Our findings provide theoretical evidence and development suggestions for future scientific governance of HCAI education. Also, the ranking of $MP > MR > AR > PSR > TR > PR > BR > MC$ reflects the key risk factors that need to be paid attention to at the present stage. Moreover, the risk framework not only systematically considers the risk governance order of HCAI education, but more importantly, it is the key bridge to the collaborative advancement of stakeholders such as managers, teachers, students, and parents in AIED applications. For example, at the procurement stage, it can provide managers with judgmental evidence on the access regulations and application safety of AIED products. At the design stage, it can provide key scaffolding and intervention directions for teachers to carry out AIED activities. In the application stage, it can provide guidance and support for students' scientific cognition and rational use of AIED tools. For parents in the promotion stage, it can help them further rationally accept AIED applications and enhance the value effect of intelligent efficiency.

7. Limitations and future works

Although this study does propose some valuable risk governance factors and potential intervention directions in HCAI education, there are still some limitations. First, our research sample used only English language articles. However, as AIED applications are being promoted and explored worldwide, publications in other languages should also be considered in future research. Moreover, the initial keywords search is limited to the two domains of HCAI and risk, which may lead to the latest AI technology reports are not being included in this study, future more studies may consider optimizing the search strategy, such as extending keywords like HCAI challenges and HCAI governance. Additionally, although the study provides a systematic risk governance framework, the current research results still lack inclusiveness, thus future analysis could go back further in time to explore the phased trends in risk governance.

In the future, if the AIED applications early warning systems can be developed according to the risk framework and index weight, it will promote the scientific, healthy and sustainable HCAI education. However, the research on effect size of each risk is lacking, especially how to provide corresponding intervention scaffolds based on the effect size. A possible future direction could be to conduct a series of meta-analyses on the specific effect sizes of each risk, so as to explore the dynamic trends and key dilemmas of risk governance in HCAI education. Another potential direction is to implement the risk framework, for example, we can carry out intervention experiments for learners in different regions, learning segments, and queues, so as to generate different types and different characteristics of avoidance strategies and promotion measures. Furthermore, attention should reach beyond AIED applications to the latest trends of HCAI education, future more studies may consider comparing the characteristics of different regions and carrying out innovative practices in HCAI education, for example, developing an index framework of the HCAI education, promoting HCAI education based on social experiments, and using multi-agent simulation experiments to simulate the trend of HCAI education.

Acknowledgement

This study is supported by the Major Projects of the National Social Science Foundation of China (grant number 19ZDA364).

References

- Ahn, J., Campos, F., Nguyen, H., Hays, M., & Morrison, J. (2021). Co-designing for privacy, transparency, and trust in K-12 learning analytics. *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 55-65). <https://doi.org/10.1145/3448139.3448145>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. In *Ethics of Data and Analytics* (pp. 254-264). Auerbach Publications. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Auernhammer, J. (2020). Human-centered AI: The Role of human-centered design research in the development of AI. *DRS International Conference 2020* (pp. 1315-1333). <https://doi.org/10.21606/drs.2020.282>
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.
- Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1-12.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.
- Cao, M., Li, D. & Wang, J. (2020). A Study of college English culture intelligence-aided teaching system and teaching pattern. *English Language Teaching*, 13(3), 77-83.
- Capatosto, K. (2017). *Foretelling the future: A Critical perspective on the use of predictive analytics in child welfare*. Ohio State University.
- Chai, C. S., Wang, X., & Xu, C. (2020). An Extended theory of planned behavior for the modelling of Chinese secondary school students’ intention to learn artificial intelligence. *Mathematics*, 8(11), 2089-2106.
- Chang, H. Y., Hsu, Y. S., & Wu, H. K. (2014). A Comparison study of augmented reality versus interactive simulation technology to support student learning of a socio-scientific issue. *Interactive Learning Environments*, 24(6), 1148–1161.

- Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: A Quantitative analysis using structural equation modelling. *Education and Information Technologies*, 25(5), 3443-3463.
- Chen, B., Hwang, G. H., & Wang, S. H. (2021). Gender differences in cognitive load when applying game-based learning with intelligent robots. *Educational Technology & Society*, 24(3), 102-115.
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, Article 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- Connor, M., & Siegrist, M. (2010). Factors influencing people's acceptance of gene technology: The Role of knowledge, health expectations, naturalness, and social trust. *Science Communication*, 32(4), 514-538.
- Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge.
- Cui, D., & Wu, F. (2021). The Influence of media use on public perceptions of artificial intelligence in China: Evidence from an online survey. *Information Development*, 37(1), 45-57.
- Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerd, J. (2021). A Review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162, Article 104094. <https://doi.org/10.1016/j.compedu.2020.104094>
- Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer. <https://doi.org/10.1007/978-3-030-30371-6>
- Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 40-60. <https://doi.org/10.17351/ests2019.260>
- Etikan, I., & Bala, K. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, 5(6), 215-217.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.
- Friedman, B., Hendry, D., & Borning, A. (2017). A Survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2), 63-125.
- Goodman, C. M. (1987). The Delphi technique: A Critique. *Journal of Advanced Nursing*, 12(6), 729-734.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42.
- Guilherme, A. (2019). AI and education: The Importance of teacher and student relations. *AI & Society*, 34(1), 47-54.
- Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program, *AI Magazine*, 40(2), 44-58.
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. The Center for Curriculum Redesign.
- Honeycutt, D., Nourani, M., & Ragan, E. (2020). Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1), 63-72.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence*, 1, Article 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- Intel Corporation. (2018). *Intel's AI privacy policy white paper: Protecting individuals' privacy and data in the artificial intelligence world*. <https://blogs.intel.com/policy/files/2018/10/Intels-AI-Privacy-Policy-White-Paper-2018.pdf>
- Jim, C. K., & Chang, H. C. (2018). The Current state of data governance in higher education. *Proceedings of the Association for Information Science and Technology*, 55(1), 198-206.
- Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3819-3828). <https://doi.org/10.1145/2702123.2702520>
- Ketamo, H. (2018). *Dreams and reality: How AI will change education*. <https://mgiep.unesco.org/article/dreams-and-reality-how-ai-will-change-education>
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14-29.
- Knox, J., Wang, Y., & Gallagher, M. (2019). *Artificial intelligence and inclusive education*. Springer.

- Kusner, M. J., & Loftus, J. R. (2020). The Long road to fairer algorithms. *Nature*, 578(2), 34-36.
- Lu, O. H. T., Huang, A. Y. Q., Tsai, D. C. L., & Yang, S. J. H. (2021). Expert-authored and machine-generated short-answer questions for assessing students' learning performance. *Educational Technology & Society*, 24(3), 159-173.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21.
- Nathanson, L., Cotcoran, S., Baker, S. C. (2013). *High school choice in New York city: A Report on the school choices and placements of low-achieving students*. Research Alliance for New York City Schools.
- Organisation for Economic Co-operation and Development (OECD). (2016). *Innovating education and educating for innovation: The Power of digital technologies and skills*. <http://dx.doi.org/10.1787/9789264265097-en>
- Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by artificial intelligence practitioners. *Information, Communication & Society*, 23(5), 719-735.
- Ozga, J. (2016). Trust in numbers? Digital education governance and the inspection process. *European Educational Research Journal*, 15(1), 69-81.
- Pasquale, F. (2015). *The Black box society*. Harvard University Press.
- Perikos, I., Grivokostopoulou, F., & Hatzilygeroudis, I. (2017). Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic. *International Journal of Artificial Intelligence in Education*, 27(3), 475-514.
- Reddy, E., Cakici, B., & Ballesterio, A. (2019). Beyond mystery: Putting algorithmic accountability in context. *Big Data & Society*, 1-7. <https://doi.org/10.1177/2053951719826856>
- Renn, O. (2021). New challenges for risk analysis: Systemic risks. *Journal of Risk Research*, 24(1), 127-133.
- Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11(5), 5-16.
- Saaty, R. W. (1987). The Analytic hierarchy process-What it is and how it is used. *Mathematical Modelling*, 9(3-5), 161-176.
- Saneiro, M., Santos, O. C., Salmeron-Majadas, S., & Boticario, J. G. (2014). Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *The Scientific World Journal*, 4, Article 484873. <http://dx.doi.org/10.1155/2014/484873>
- Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29(2), 353-400.
- Schmidt, A. (2020). Interactive human centered artificial intelligence: A Definition and research challenges. In *Proceedings of the International Conference on Advanced Visual Interfaces* (pp. 1-4). <https://doi.org/10.1145/3399715.3400873>
- Schweizer, P. J. (2021). Systemic risks—Concepts and challenges for risk governance. *Journal of Risk Research*, 24(1), 78-93.
- Sharples, M. (2019). *To improve education: Focus on pedagogy not technology*. <https://oeb.global/oeb-insights/to-improve-education-focus-on-pedagogy-not-technology>
- Shneiderman, B. (2020a). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504.
- Shneiderman, B. (2020b). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109-124.
- Sivill, T. (2019). Ethical and statistical considerations in models of moral judgments. *Frontiers in Robotics and AI*, 6, Article 39. <https://doi.org/10.3389/frobt.2019.00039>
- Turón, A., Aguarón, J., Escobar, M. T., & Moreno-Jiménez, J. M. (2019). A Decision support system and visualisation tools for AHP-GDM. *International Journal of Decision Support System Technology (IJDSST)*, 11(1), 1-19.
- The United Nations Educational, Scientific and Cultural Organization (UNESCO). (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development*. [http://www.nied.edu.na/assets/documents/05Policies/NationalCurriculumGuide/Artificial_Intelligence_\(AI\)-challenges_and_opportunities_for_sustainable_development.pdf](http://www.nied.edu.na/assets/documents/05Policies/NationalCurriculumGuide/Artificial_Intelligence_(AI)-challenges_and_opportunities_for_sustainable_development.pdf)
- Verkijika, S. F., & De Wet, L. (2015). Using a brain-computer interface (BCI) in reducing math anxiety: Evidence from south Africa. *Computers & Education*, 81(2), 113-122.
- Wang, Y., Liu, C., Tu, Y.F. (2021). Factors affecting the adoption of AI-based applications in higher education: An Analysis of teachers' perspectives using structural equation modeling. *Educational Technology & Society*, 24(3), 116-129.

- Wessels, B. (2015). Authentication, status, and power in a digitally organized society. *International Journal of Communication*, 9(1), 2801–2818.
- White, J. M., & Lidskog, R. (2022). Ignorance and the regulation of artificial intelligence. *Journal of Risk Research*, 25(4), 488-500.
- Winters, N., Eynon, R., Geniets, A., Robson, J., & Kahn, K (2020). Can we avoid digital structural violence in future learning systems? *Learning, Media and Technology*, 45(3), 17-30.
- Wu, W., Huang, T., & Gong, K. (2020). Ethical principles and governance technology development of AI in China. *Engineering*, 6(3), 302-309.
- Xu, W. (2019). Toward human-centered AI: A Perspective from human-computer interaction. *Interactions*, 26(4), 42-46.
- Xu, W., Ge, L. Z., & Gao, Z. F. (2021). Human-AI interaction: An Emerging interdisciplinary domain for enabling human-centered AI. *CAAI Transactions on Intelligent Systems*, 16(4), 605–621.
- Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers & Education: Artificial Intelligence*, 2, Article 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1-27.
- Zhang, J. (2021). Reform and innovation of artificial intelligence technology for information service in university physical education. *Journal of Intelligent & Fuzzy Systems*, 40(2), 3325-3335.
- Zhang, Y., Qin, G., Cheng, L., Marimuthu, K., & Kumar, B. S. (2021). Interactive smart educational system using AI for students in the higher education platform. *Journal of Multiple-Valued Logic & Soft Computing*, 36(1-3), 83-98.

AI, Please Help Me Choose a Course: Building a Personalized Hybrid Course Recommendation System to Assist Students in Choosing Courses Adaptively

Hui-Tzu Chang¹, Chia-Yu Lin^{2*}, Wei-Bin Jheng³, Shih-Hsu Chen⁴, Hsien-Hua Wu⁵, Fang-Ching Tseng⁶ and Li-Chun Wang⁴

¹Center for Institutional Research and Data Analytics, National Yang Ming Chiao Tung University, Hsinchu //

²Department of Computer Science and Information Engineering, National Central University, Taoyuan //

³Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu // ⁴Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu // ⁵Department of Transportation and Logistics Management, National Yang Ming Chiao Tung University, Hsinchu // ⁶Electrical Engineering and Computer Science Undergraduate Honors Program, National Yang Ming Chiao Tung University, Hsinchu //

simple@nycu.edu.tw // sallylin0121@gmail.com // sozai97.cs06@nctu.edu.tw // brian880403@gmail.com // simonwu.mg09@nycu.edu.tw // claire.eecs07@nctu.edu.tw // wang@nycu.edu.tw

*Corresponding author

ABSTRACT: The objective of this research is based on human-centered AI in education to develop a personalized hybrid course recommendation system (PHCRS) to assist students with course selection decisions from different departments. The system integrates three recommendation methods, item-based, user-based and content-based filtering, and then optimizes the weights of the parameters by using a genetic algorithm to enhance the prediction accuracy. First, we collect the course syllabi and tag each course from twelve departments for the academic years of 2015 to 2020. Next, we use the course tags, student course selection records and grades to train the recommendation model. To evaluate the prediction accuracy, we conduct an experiment on 1490 different courses selected by 5662 students from the twelve departments and then use the root-mean-squared error and the normalized discounted cumulative gain. The results show that the influence of item-based filtering on the course recommendation results is higher than that of user- and content-based filtering, and the genetic algorithm can find the optimal solution and the corresponding parameter settings. We also invite 61 undergraduate students to test our system, complete a questionnaire and provide their grades. Overall, 83.60% of students are more interested in courses at the top of the recommendation lists. The students are more autonomously motivated rather than holding extrinsic informational motivation across the hybrid recommendation method. Finally, we conclude that PHCRS can be applied to all students by tuning the optimal weights for each course selection factor for each department, providing the best course combinations for students' reference.

Keywords: Human-centered AI in education, AI course recommendation system, Learning aids in systems

1. Introduction

In recent years, the number of research works applying Artificial Intelligence (AI) to educational systems have increased rapidly. AI offered a new solution for education as it helped develop an adaptable, inclusive, agile, individualized, and effective learning environment to overcome the disadvantages of traditional education or training. Additionally, it also brought hope and potential of innovation for education (Renz et al., 2020; Renz & Vladova, 2021). In those AI systems, human-centered AI in education enables us to gain a deeper understanding of students' learning behaviors, reaction time, emotion, or needs (Renz et al., 2020; Yang et al., 2021). It also helped students find the potential and problems, then set up study plans for them using information and communications technologies (ICTs; Yang et al., 2021). A system that assists students with planning their courses was extremely important (Lin et al., 2018). Recent course recommendation system research has focused on how to precisely recommend students courses that suit their needs, with many works proposing course selection methods and algorithms to deal with course recommendation, though none of the methods were designed based on human-centered AI in education. The focus of these studies was on raising the grades of the students (Chang et al., 2016), their graduation rate (Kurniadi et al., 2019) or their employment rate (Farzan & Brusilovsky, 2006) rather than the personal factors affecting the recommendation process.

Course options are important for students to fulfill their degree requirements and to determine their future career directions (Farzan & Brusilovsky, 2006; Kurniadi et al., 2019). In response to the trend of higher education, institutions promote interdisciplinary courses and distance learning courses, and AI systems to contribute to the selection of courses with more diversity (Chang & Chen, 2021). When students are faced with information overload as they are selecting courses, students' adaptive development would be secured if their school provided

a course recommendation system that recommended courses based on their interests, abilities, and career goals (Iatrellis et al., 2017; Sawarkar et al., 2018). Thus, this study proposes the personalized hybrid course recommendation system (PHCRS) that considers students' course selection factors to provide better course selection advice. PHCRS utilizes the course selection data (e.g., courses, grades) and course data (e.g., objectives, knowledge area, skills) accumulated by the school for system development. To ensure that these factors can help generate a better recommendation result, this study uses a genetic algorithm to determine the importance of each indicator and recommendation method, applies weights to the recommendation process, and provides advice to students.

2. Literature review

2.1. Human-centered AI in education

Previous AI technologies focused on how to behave and think like a human, while recent research switched their focus to human-centered AI (HCAI), a technology that approaches AI from a human perspective through human environments (Renz & Vladova, 2021; Yang et al., 2021). Human-centered AI needs explainable computation and decision-making processes, social phenomena, and mankind characteristics to adjust its algorithms to help enhance human intelligence using machine learning to increase human welfare (Yang et al., 2021). HCAI has been applied to a wide variety of domains, and its effectiveness in education is of great importance. In addition, HCAI can help students learn, adapt, integrate, self-correct, and use data to tackle complicated tasks in the hopes of solving more learning, emotion or career development problems that students may face. AI is superior to humans when it comes to computing and decision making, and it can also educate and train humans to enhance their performances, as well as mine implicit values (Yang et al., 2021). With the development of AI, the trend of education has shifted from the one-size-fits-all approach to the precision approach (Zawacki-Richter et al., 2019), which utilizes AI for analysis. The precision approach identifies students in need and offers real-time assistance, which enhances the teaching quality and learning outcomes for students. It also enables students to develop their skills and knowledge in a more personalized way by providing more precise information, understanding the students' progress of, and what should be done to realize their goals (Yang et al., 2021).

Even though more and more services offer data-driven smart learning solutions for education, only a small portion of them apply AI techniques (Liu et al., 2023). Ahmad et al. (2020) reviewed previous research on applications of AI in education and split the domains into intelligent tutoring systems (ITS), evaluation, adaptive learning, recommendation systems, student performance, sentiment analysis, detention or drop out, and course monitoring. Among those topics, ITS is the most popular and the most important because it allows teachers to provide adaptive learning routes in educational environments and assist students with planning their own learning routes based on their personal interests, abilities, or future career development (Alkhatlan & Kalita, 2019). Even though AI has a lot of potential if applied in education and is increasingly gaining popularity, only few are implementing AI in education tools and even less of them use these tools in their institutions; thus we can conclude that people still doubt the ability or reliability of AI, which limits the development of HCAI. More research has advocated not use AI to replace humans (Xu, 2019), but to support humans based on human's benefit (Schmidt, 2020). Education relatives have come to an understanding that the use of HCAI is to help realize the goals of positive learning outcomes and teaching success instead of replacing traditional education methods, then diminish the fear of AI from students and teachers afterwards (Renz & Vladova, 2021).

2.2. AI recommendation systems in education

ICTs play a huge role in the globalization era and information society, while also providing new opportunities for many domains. In education, ICTs are utilized for the teaching and learning process (Urdaneta-Ponte et al., 2021). However, the development of ICTs poses some challenges, including the increasing complexity and loading of information can make students spend too much time on searching for information and consumes the amount of time they are able to spend studying, which would decrease and their grades would decrease accordingly. If students can get reliable and adequate information easier and quicker, it would be a decisive factor in their learning outcomes. To resolve this problem, the course recommendation system is developed, and the goal of the course recommendation system is to offer choices and recommendations for each student based on their needs, helping students find the courses that truly meet their requirements through information filtering, data mining and predictive algorithms.

The main approaches used in course recommendation system are the collaborative filtering, content-based filtering, and hybrid recommendation methods (Urdaneta-Ponte et al., 2021). (1) Collaborative filtering. There are two main filtering methods, including item-based and user-based filtering. Item-based filtering uses students' grades in other subjects or domains to provide course recommendations (Dwivedi & Roshni VS, 2017). User-based filtering matches the course selection route history of a current student to an alumnus who shared a similar route, then recommends the course list of the alumnus to the current student (Zhang et al., 2015). (2) Content-based filtering. The filtering mechanism is built upon the characteristics of the course syllabi, such as the subject field or the lecture content, thus providing a course list similar to one's interested subjects or domains (Esteban et al., 2020). However, these methods have their respective strengths and weaknesses; to address the disadvantages of the methods mentioned above, researchers have proposed (3) hybrid recommendation methods (Çano & Morisio, 2017). The collaborative filtering and content-based filtering hybrid recommendation method is the most common method since it overcomes the limitations of both filtering methods above, increases predictability, and decreases the degree of sparsity and the loss of information (Esteban et al., 2020).

Several AI technologies are introduced for the construction of the course recommendation system in recent years, including Bayesian techniques, artificial neural networks, machine learning techniques, genetic algorithms, and fuzzy set techniques. These AI techniques prove to be adequate for designing recommendation systems in the big data era (Urdaneta-Ponte et al., 2021), and a genetic algorithm is one of the most often used method. A genetic algorithm, proposed by Holland (1975), was inspired by the encoding and decoding process of DNA and applied to the artificial environment. A genetic algorithm can automatically optimize the weights of each criterion and variable in the recommendation system through the optimization of likelihood function (Esteban et al., 2020) to obtain the final estimation for the system (Esteban et al., 2020). However, even though a genetic algorithm has shown good performance when used in building recommendation systems, only research applies this method (Esteban et al., 2020). Esteban et al. (2020) used hybrid filtering combining collaborative filtering and content-based filtering to train a course recommendation model, then applied a genetic algorithm to optimize the weights of student information, course information, recommendation methods, and system attributes to build a course recommendation system with high accuracy for students. A genetic algorithm has also been applied to estimate the best learning path. Dwivedi and Roshni (2017) matched the learning path of current students with alumni history data and then used a genetic algorithm to find the best learning path for each current student. Huang et al. (2007) applied computerized adaptive testing combined with a genetic algorithm and case-based reasoning to build the best learning path of online courses. In conclusion, a genetic algorithm is a useful tool in learning systems; it provides the best solution for complicated problems that students encounter, and its computation results can also be a reference for students' course selection and learning path.

For the reasons mentioned above, we propose PHCRS for formal offline courses to consider the different learning needs of students. The system offers a course recommendation list based on personalized course selection factors, decision sequences and course importance to satisfy the personalized study, capacity building and career exploration needs of students. To achieve this goal, we first filter the factors affecting the students' course selection decisions as the indicators of system development and then use a hybrid multicriteria recommendation method to develop the recommendation system. Last, we use a genetic algorithm to find the weights of student information, course information and system attributes with the goal of determining weights in a standardized manner and optimizing system attributes automatically. This study proposes three hypotheses to verify the effectiveness of PHCRS.

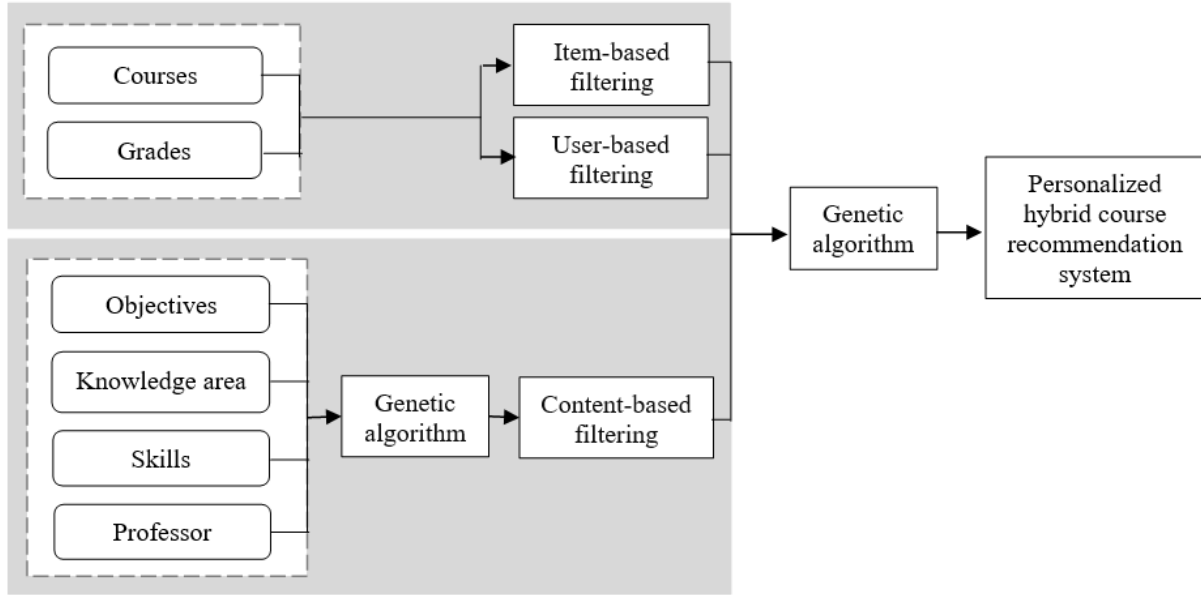
- Hypothesis 1: Students' degree of interest in the courses recommended by the hybrid recommendation method will differ among the course recommendation order.
- Hypothesis 2: The degree of interest in the courses recommended to a student will be affected by the student's internal and external motivations for taking a course.
- Hypothesis 3: Students' degree of academic performance in the courses recommended will differ among those following and not following the recommendation list.

3. Development of a personalized hybrid course recommendation system

The steps of the research design process are shown in Figure 1. We first transform and encode the data used for system development and then apply item-based, user-based and content-based filtering to compute information regarding students and courses to obtain the results of each recommendation method. Then, we use a genetic algorithm to automatically optimize the weights for all filtering methods, and the optimal parameter settings for each student can be found, thus achieving the effect of adaptive recommendation. Finally, we use the root-mean-

squared error (RMSE) and normalized discounted cumulative gain (NDCG) to evaluate the effectiveness of the system, thus forming PHCRS. The detailed process is discussed in the following sections.

Figure 1. A Framework for PHCRS



3.1. Data description and preparation

We used course and student data from the Center for Institutional Research and Data Analytics at National Yang Ming Chiao Tung University (NYCU) to train the recommendation system (see Table 1). These data included 12 departments from the Colleges of Electrical and Computer Engineering, Computer Science, Engineering, Management, and Hakka Studies, and a total of 6766 courses were provided from the fall 2015 semester to the fall 2020 semester. For student information, a total of 5662 students from the 12 departments who were enrolled between 2015 and 2020 were selected. To prepare the training data, the researchers collected the course outlines and interviewed the teachers via telephone. The two researchers in each department discussed and agreed upon the labeling rules and then compared the similarities and differences in the labeling results after making the labels. In cases of disagreement, the scorers discussed the issue until a consensus was reached. The interrater reliability was between .7 and .8. The attributes of each course were labeled as follows: (1) Course objectives: This label indicates what the course mainly teaches students, such as signal processing or communication systems. There is a total of 377 possible labels from the 12 departments. (2) Knowledge areas: This label is based on the theories, methods or empirical theories from the field of electrical engineering that are taught to students, such as information and communication, system-on-chip, and 126 other areas from the 12 departments. (3) Skills: This label is based on the relevant technologies, resources or tools used in each course, such as Python or MOSFET. There are a total of 1744 possible labels from the 12 departments. (4) Professors: This label indicates who the course instructor is. After the data preparation, three recommendation methods and a genetic algorithm optimization are implemented in PHCRS for students with different learning needs, as shown below.

Table 1. Student and course information

College/Department	Students	Courses	Total number of courses	Label			
				Course objectives	Knowledge areas	Skills	Professor
College of Electrical and Computer Engineering							
• Department of Electrical and Computer Engineering	1258	332	1890	45	15	373	188
• Department of Photonics	209	94	369	19	3	76	47
College of Computer Science							
• Department of Computer Science	1171	235	1013	54	7	306	114
College of Engineering							
• Department of Civil	473	122	688	70	6	103	51

Engineering							
• Department of Mechanical Engineering	596	119	671	47	11	87	58
• Department of Materials Science and Engineering	299	74	367	24	8	68	37
College of Management							
• Department of Management Science	285	72	235	11	11	204	23
• Department of Transportation & Logistics Management	280	70	336	10	2	81	23
• Department of Industrial Engineering and Management	314	68	256	9	20	149	21
• Department of Information Management and Finance	268	64	283	10	3	163	29
College of Hakka Studies							
• Department of Humanities and Social Sciences	268	132	370	21	5	52	29
• Department of Communication and Technology	241	108	288	57	35	82	26
Total	5662	1490	6766	377	126	1744	646

Note. #in academic years 104 to 109.

3.2. Recommendation model construction

The method we used for recommendation is a multicriteria hybrid recommendation method integrating item-based, user-based and content-based filtering. The formula for predicting the score that *student i* gives to *course j* is as follows: $p_{ij} = \alpha \cdot ICF_{ij} + \beta \cdot UCF_{ij} + \gamma \cdot CBF_{ij}$ (1), where $\alpha + \beta + \gamma = 1$, ICF_{ij} is the score that *student i* gives to *course j* based on item-based filtering, UCF_{ij} is the score that *student i* gives to *course j* based on user-based filtering, and CBF_{ij} is the score that *student i* gives to *course j* based on content-based filtering. The range of the predicted scores of all methods is between 1 and 4.3.

Item-based filtering: Item-based collaborative filtering calculates the similarity score between courses and recommends similar courses (Sarwar et al., 2001). We find the students who have taken the two courses and calculate the difference of their scores in the two courses. The smaller the difference is, the higher the similarity. The similarity is represented as $w_{i,j}$ and is shown in (2), where A is the set of students who have taken *course i* and *course j*. Assuming *student x* has taken *course i*, if PHCRS wants to recommend *course k* to *student x*, the predicted score is calculated by formula (3). The numerator is equal to the product of $w_{i,k}$ and the student's grade in *course i*. The denominator is the summation of the similarity between *course i* and *course k*.

$$\text{Similarity between course } i \text{ and course } j (w_{i,j}) = \frac{1}{1 + \sqrt{\sum_{A \in M(i) \cap M(j)} (\text{grade}(A,i) - \text{grade}(A,j))^2}} \quad (2)$$

$$\text{Prediction score of course } k \text{ for student } x = \frac{\sum_{w_{i,k} > 0} \text{grade}(x,i) * w_{i,k}}{\sum_{w_{i,k} > 0} w_{i,k}} \quad (3)$$

User-based filtering: User-based collaborative filtering utilizes students' past course data to calculate the similarity between students and recommend courses taken by similar students (Han et al., 2016). To calculate the similarity between two students, we have to determine the courses the students have both taken. We utilize the scores of two students in the courses to calculate the similarity. The similarity of *student x* and *student y* is represented as a weighted value ($w_{x,y}$) as shown in (4), where $N(x)$ are the courses that *student x* has taken and $N(y)$ are the courses that *student y* has taken. If the scores are closer, the similarity of the two students is higher. If PHCRS wants to recommend *course k* to *student y*, the similarity of *student x* and *student y* is multiplied by the scores of *student x* on *course k*. The average weighted value is the predicted score, as shown in (5).

$$\text{Similarity of student } x \text{ and student } y (w_{x,y}) = \frac{1}{1 + \sqrt{\sum_{i \in N(x) \cap N(y)} (\text{grade}(x,i) - \text{grade}(y,i))^2}} \quad (4)$$

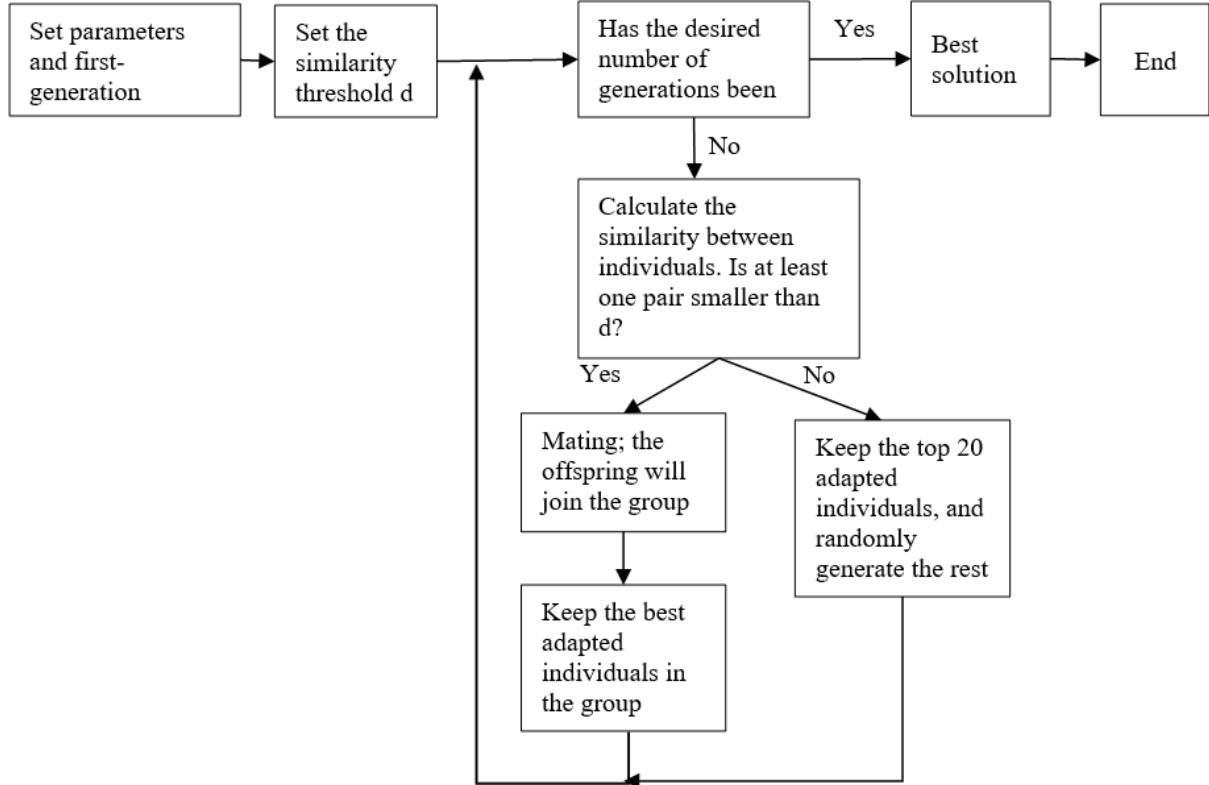
$$\text{Predicted score for student } y \text{ on course } k = \frac{\sum_{w_{x,y} > 0.2} \text{grade}(y,k) * w_{x,y}}{\sum_{w_{x,y} > 0.2} w_{x,y}} \quad (5)$$

Content-based filtering: Content-based filtering recommends similar courses based on the characteristics of students' past courses (Esteban et al., 2020). In the first step, the feature vectors of the courses are extracted. The course feature vector indicates which domains the courses belong to and which objectives the courses contain. To calculate the feature vectors of *student x* for *course i*, the feature vector of *course i* is multiplied by the score of *student x* on *course i*. We add up all the feature vectors of *student x* on each course and define this value as the feature vector of *student x*. To recommend *course j* to *student x*, we use the feature vector of *student x* and the feature vector of *course j* to calculate the cosine value ($\cos\theta = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \cdot \|\vec{j}\|}$) as the similarity. If the similarity is close to 1, *student x* is more likely to like *course j*.

3.3. Weight selection

We apply a genetic algorithm to find the optimal solution for course recommendation. The genetic algorithm is a type of machine learning algorithm that finds new and better individuals through crossover or mutation of candidate individuals; this procedure iterates for multiple generations until the ending criteria are satisfied (Holland, 1975). The ending criterion in this study is a fixed number of evolutions. Our algorithm follows the algorithm proposed by Esteban et al. (2020). The flow chart of the genetic algorithm in computing the optimum solution is shown in Figure 2, and the details of each step are explained in the next section.

Figure 2. Flowchart of the proposed genetic algorithm



Each individual has eleven genes and is split into four parts (Figure 3), where z_i represents the i^{th} gene.

Figure 3. Gene paradigm

7	1	76	83	1	60	55	100	1	100	2
---	---	----	----	---	----	----	-----	---	-----	---

The first three genes represent the weights of item-based, user-based and content-based filtering, respectively, when combining their solutions. In other words, $\alpha = z_1/(z_1 + z_2 + z_3)$, $\beta = z_2/(z_1 + z_2 + z_3)$ and $\gamma = z_3/(z_1 + z_2 + z_3)$. For example, if $z_1 = 7$, $z_2 = 1$ and $z_3 = 76$, then $\alpha = \frac{7}{7+1+76} = 0.083$, $\beta = \frac{1}{7+1+76} = 0.011$ and $\gamma = \frac{76}{7+1+76} = 0.904$.

The fourth to seventh genes represent the weights of content-based filtering for each variable, which are used to calculate the similarity between students. The variables include the domain of the course, the overview of the course and the detailed course context and lecturer, represented by α , β , γ and δ . For example, if $z_4 = 83$, $z_5 = 1$, $z_6 = 60$ and $z_7 = 55$, then $\alpha = \frac{83}{83+1+60+55} = 0.417$, $\beta = \frac{1}{83+1+60+55} = 0.005$, $\gamma = \frac{60}{83+1+60+55} = 0.301$ and $\delta = \frac{55}{83+1+60+55} = 0.276$.

The eighth and ninth genes represent the weights of user-based filtering for each variable, which are used to calculate the similarity between students, where z_8 is always 100. For example, $z_9 = 1$ means that the threshold of user-based filtering is 0.1.

The tenth and eleventh genes represent the weights of item-based filtering for each variable, which are used to calculate the similarity between students, where z_{10} is always 100. For example, $z_{11} = 2$ means that the threshold of item-based filtering is 0.2.

3.4. Parameters in a genetic algorithm

The following sections introduce different formulas for the genetic algorithm that were designed.

3.4.1. Distance threshold d

To address the inability of highly similar existing individuals to generate a different child generation and find the optimal solution, the generation process restarts when two genes of a child generation are too similar. The similarity threshold of *distance* d is set to 0.8. If the similarity between every individual pair is higher than d , then the process enters the “restart” phase, meaning that the 20 best individuals are kept while the others are generated randomly.

3.4.2. Individual dissimilarity

We use the Hamming distance to calculate the distance between each pair of individuals and then transform the distance into a similarity value, which is the number of genes that are the same divided by the length of the individual ($L = 11$). For example, when the first, third and fourth genes in a pair of individuals are the same, the similarity is $\frac{3}{11} \approx 0.27$.

3.4.3. Crossover operator

The method of generating a child generation is to cross the same set of genes from two parent generation individuals. For example, the first and third genes of a child-generation individual may be from the father, and the second and fourth genes may be from the mother. The crossover probability of a set of genes is 50%.

3.4.4. Update process

The best individuals of each child generation are kept, maintaining the total number of individuals, and then the next generation is generated.

3.5. Evaluation metrics

This study uses the Root Mean Square Error (RMSE) and the Normalized Discounted Cumulative Gain (NDCG) to evaluate the recommendation results.

RMSE: The RMSE has been used as a standard statistical metric to measure model performance in the recommendation system (Esteban et al., 2020). When there are more samples or the error distribution is expected to be Gaussian, reconstructing the error distribution using RMSEs will be even more reliable (Chai & Draxler, 2014). The purpose of the RMSE is to compare the predicted score of *student i* for *course j*, v_{ij} , and the real

score given by the student, v_{ij} . For the testing data set $K=\{(i,j)\}$, $RMSE = \sqrt{\frac{\sum_{(i,j) \in K} \sum (p_{ij} - v_{ij})^2}{\#K}}$, and a smaller RMSE value means that the predicted score is closer to the real score given by the student.

NDCG: The NDCG is a family of ranking measures widely used in applications. It has two advantages. First, the NDCG allows each retrieved document has graded relevance while most traditional ranking measures only allow binary relevance. Second, the NDCG involves a discount function over the rank, while many other measures uniformly weight all positions (Wang et al., 2013). For the k example courses, we sort the courses by the recommendation scores and calculate the discounted cumulative gain (DCG). The DCG is shown in (5), where k represents the number of courses the system recommends and rel_i is the gain for each recommended course. In the evaluation, when the recommended course overlaps the real record, we set the gain rel_i to 1; otherwise, it is set to 0. The ideal course order based on the predicted score is used to calculate the ideal discounted cumulative gain (IDCG), as shown in (6). We can use the DCG and IDCG to calculate the NDCG, as shown in (7).

$$DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (5)$$

$$IDCG_k = \sum_{i=1}^{|rel_k|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (6)$$

$$NDCG_k = \frac{DCG_k}{IDCG_k} \quad (7)$$

3.6. Experimental work

The experiment is divided into two parts. First, we determine the optimized weight for each index in PHCRS (including item-based filtering, user-based filtering, and content-based filtering) separately. Then, we use the RMSE and NDCG to evaluate the accuracy of the recommendation provided by PHCRS. The system is built in the Python environment, including the recommendation criterion, genetic algorithm, and system performance evaluation. The data source is the course selection records of college students from twelve departments at NYCU from academic years 2015-2020. The unit of the experiment during system development is per department, the training data consist of the course selection data from 2015-2018 and the 2020 academic year, and the testing data are the course selection data from 2019. The results of the experiment are given below.

3.6.1. Criteria weight optimization

The first part of the experiment uses a genetic algorithm to determine the weights of the three recommendation methods of PHCRS, to optimize their relative parameters, and to evaluate the influence of the weights on PHCRS. In PHCRS, there are nine weights that need to be optimized, including the weights of item-based, user-based and content-based filtering, the sizes of the filters of item-based filtering and user-based filtering, and the weights of the objectives, knowledge areas, skills and professors in content-based filtering. The settings of the important parameters of the genetic algorithm are shown in Table 2, which we applied for the experiment.

Table 2. Configuration of the genetic algorithm parameters

Parameter	Value
Number of generations	100
Population size	209-1258
Crossover probability	0.9
Initial value for incest prevention threshold	4
Allowed range for weight genes	[0, 50]
Allowed range for neighborhood gene	[1, 50]
Allowed range for metric genes	[0, 4] or [0, 1]

Table 3 shows the optimized weights of each department obtained by the genetic algorithm. The results showed that there is a large difference between the weights of the four indexes in content-based filtering, with the weight of “Course objectives” lying within .339% ~ 78.723%, the weights of “Knowledge areas” lying within 1.613% ~ 38.525%, the weights of “Skills” lying within .633% ~ 38.672%, and the weights of “Professor” lying within 7.447% ~ 53.714%, indicating that the influence of the indexes differs from department to department. For example, students from the Department of Electrical and Computer Engineering mainly consider “Professor” (53.459%), and students from the Department of Mechanical Engineering mainly consider “Course objectives.” We further compare the weights of the three recommendation methods in PHCRS, and the results show that for all departments, the weight of item-based filtering is always the highest, lying within 94.118% ~ 98.039%, while the weights of user-based and content-based filtering are both low in PHCRS; the former lies within .971% ~ 5.208%, and the latter lies within .908% ~ 2.913%. Thus, item-based filtering is the method that mainly influences the results of course recommendation provided by PHCRS.

Table 3. Criteria weights, similarity measures chosen by genetic algorithm, and RS evaluation

College/ Department	Content-based filtering				Hybrid recommendation			Evaluation	
	Course objectives	Knowledge areas	Skills	Professor	Item-based filtering	User-based filtering	Content-based filtering	RMSE	NDCG
College of Electrical and Computer Engineering									
• Department of Electrical and Computer Engineering	.63%	13.84%	32.08%	53.46%	97.47%	1.27%	1.27%	.61	.93
• Department of Photonics	.49%	31.53%	37.93%	30.05%	96.77%	1.08%	2.15%	.37	.94
College of Computer Science									
• Department of Computer Science	17.62%	31.09%	19.69%	31.61%	96.15%	2.56%	1.28%	.90	.90
College of Engineering									
• Department of Civil Engineering	.41%	38.53%	25.00%	36.07%	95.89%	2.74%	1.37%	.80	.93
• Department of Mechanical Engineering	78.72%	6.38%	7.45%	7.45%	95.83%	3.13%	1.04%	.58	.95
• Department of Materials Science and Engineering	.34%	33.22%	33.90%	32.54%	94.12%	4.90%	.98%	.56	.96
College of Management									
• Department of Management Science	49.37%	32.28%	.63%	17.72%	98.04%	.98%	.98%	.42	.96
• Department of Transportation & Logistics Management	10.86%	21.14%	14.29%	53.71%	95.75%	2.13%	2.13%	.59	.95
• Department of Industrial Engineering	69.36%	1.61%	8.07%	20.97%	93.75%	5.21%	1.04%	.54	.93

and Management									
• Department of Information Management and Finance	.39%	28.52%	38.67%	32.42%	96.12%	.97%	2.91%	.39	.97
College of Hakka Studies									
• Department of Humanities and Social Sciences	70.27%	8.11%	5.41%	16.22%	97.00%	1.00%	2.00%	.47	.95
• Department of Communication and Technology	29.31%	22.66%	29.31%	18.73%	97.67%	1.16%	1.16%	.75	.96

3.6.2. RS evaluation

The second part of the experiment uses the RMSE and NDCG to evaluate the accuracy of the course recommendation results provided by PHCRS. The value of RMSE indicates the difference between the predicted score and the score provided by students who finished the course. A larger RMSE value means that the difference between the predicted and real scores is larger. The results showed that the RMSE values of all departments lie within .365 ~ .898, with the departments with fewer courses having lower RMSE values (e.g., the Department of Photonics) and the departments with more courses having higher RMSE values. On the other hand, the value of NDCG indicates the sequence of recommendations, and a larger value of NDCG means that a more highly correlated course could be recommended first (e.g., courses that could yield higher grades). The results showed that the value of NDCG lies within .902 ~ .970 for all departments, meaning that for all departments, the collaborative filtering method applied by PHCRS is able to recommend courses to students based on the importance of the course (Table 3). It is worth noting that even though there is no direct relationship between the performance of RMSE and NDCG, generally, the departments with good RMSE performance also have sufficient NDCG values.

Figure 4 shows the results of the genetic algorithm iterating for 100 generations on each department. From the scree plot of each department, the RMSE values of the first generation lie within .4 ~ 2.8, and as the evolution continues, the RMSE values for every department decrease to .4 ~ .9, indicating that using a genetic algorithm in collaborative filtering can yield the optimal solution. We also find that the convergence for the College of Engineering is more obvious, and the similarity of the College of Management courses is higher, but both are able to minimize the recommendation error as evolution continues.

4. Research design

This study uses a survey method to verify the accuracy of PHCRS. The survey uses nonprobability sampling to invite undergraduates from the Department of Electrical and Computer Engineering, and Computer Science, NYCU, who volunteered as participants. As the freshmen's course selection and grade data were not yet completed, they were excluded to avoid interference in the research results. A total of 61 students were selected (28 sophomores, 15 juniors, and 18 seniors; 44 males and 17 females). In this research, recruitment posters were sent out by online student communities. After the students signed up, the researchers explained the research process and the parameters via phone or mail. To collect the data, students were required to log in to the course recommendation system. After reading the description of the hybrid recommendation method, students were asked to evaluate whether the courses recommended by the method was of interest, and if so, to provide their reasoning. Finally, they were asked to fill in their personal information and offer suggestions for the system.

This study uses a recommendation effect scale defined by our research group. When students browsed the course recommendation list, they were asked to evaluate whether each course was of interest to them and the reasons for their answer. For example, when students answered, "yes", they would select from reasons aligned with

“autonomous motivation,” which comes from careful consideration and self-determination (Lee & Sun, 2010) and includes reasons, such as the practicality of the course content, individual learning plans and personal interests. In addition, there were other reasons aligned with passive “external information motivation” (Lee & Sun, 2010), which included reasons, such as making up for missed credits, the course being easy to pass, and seeing good reviews about the teacher. This study also uses the students’ true course selection list and grade data to verify the accuracy of PHCRS.

5. Data analysis and results

5.1. An analysis of the difference among the students’ degree of interest in the courses recommended according to the order of the recommendations

Chi-Square test is used in this section. The data follow a normal distribution (skewness between -.48 and 1.30; kurtosis between -2.28 and 1.08). Table 4 shows that the course recommendation order is related to the students’ interest or not ($\chi^2 = 10.38$; $p < .05$). The results indicated that the students were more interested in the courses at the top of the recommendation lists.

Table 4. A difference analysis between the students’ degree of interest in the courses recommended in the course recommendation order

Recommendation Courses	<i>n</i>	Interest		No interest		χ^2	<i>p</i>
		Frequency	Percentage	Frequency	Percentage		
First course	61	51	83.60%	10	16.40%	10.38	.03*
Second course	61	46	75.40%	15	24.60%		
Third course	61	43	70.50%	18	29.50%		
Fourth course	48	30	62.50%	18	37.50%		
Fifth course	46	27	58.70%	19	41.30%		

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

5.2. The degree of interest in the recommended courses is affected by students’ internal and external motivations for taking a course

The Mann-Whitney U nonparametric test is used in this section. The data follow a normal distribution (skewness between .13 and 1.58; kurtosis between -1.06 and 2.10). Table 5 shows that the proportion of students with autonomous motivation ($M = 89.13\% \sim 100\%$) was higher than that of students with extrinsic informational motivation ($M = 29.63\% \sim 44.19\%$; $p < .001$) across the five recommendation courses. The results indicated that most students choose courses according to their plans, interests, or needs.

Table 5. A difference analysis of the students’ motivation of course-taking in hybrid recommendation method

Recommendation Courses	<i>n</i>	Autonomous motivation		Extrinsic informational motivation		<i>p</i>
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
First course	51	92.16%	27.15%	43.14%	50.02%	.00***
Second course	46	89.13%	31.47%	41.30%	49.78%	.00***
Third course	43	97.67%	15.25%	44.19%	50.25%	.00***
Fourth course	30	96.67%	18.26%	40.00%	49.83%	.00***
Fifth course	27	100.00%	0.00%	29.63%	46.53%	.00***

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

5.3. The degree of academic performance in the courses recommended is affected by students’ following and not following the recommendation list

The Mann-Whitney U nonparametric test is again used in this section. The data follow a normal distribution (skewness = -.34; kurtosis = .54). Table 6 shows that the students’ degree of academic performance in the courses recommended will not differ among the following ($M = 85.90$) and not following the recommendation list ($M = 84.98$; $p > .05$). The results indicated that there is same on their academic performance.

Table 6. A difference analysis between the students' degree of academic performance in the courses recommended among following and not following the recommendation list

Recommendation Courses	n	Recommendation and true course selection list overlap proportion		Academic performance				p
				Following the recommendation list		Not following the recommendation list		
		Min %	Max %	M	SD	M	SD	
Given 5 recommended courses	61	20%	100%	85.90	7.37	84.98	5.06	

6. Conclusions

This study proposes PHCRS based on human-centered AI in education combining item-based filtering, user-based filtering, and content-based filtering to recommend courses for students from different departments in universities. Our system used a genetic algorithm to automatically optimize the weights of indexes. In addition to enhance the accuracy of PHCRS, a genetic algorithm also configures the weights of different recommendation methods for each student to suit their needs. The results show that the weights of recommended methods are slightly different between departments. However, the influence of item-based filtering on the course recommendation result is higher than that of user-based and content-based filtering, meaning that students tend to select courses with similar characteristics. This result is in line with that of Chang et al. (2022), who found that the accuracy of item-based filtering is better than that of other recommendation methods through the receiver operating characteristic (ROC) curve. We also found that after the experiment, the weights of the four parameters in content-based filtering (course objectives, knowledge areas, skills, and professors) were not the same, meaning that the focus of students on courses was different.

We use RMSE and NDCG to evaluate the effectiveness of PHCRS, and the results show superior performance compared to previous research (ex: Esteban et al., 2020; Defiebre et al., 2022; Ngaffo et al., 2020). This study also collects data of university students who used PHCRS to evaluate the helpfulness of the system on students in real world situations. The results show that students are more interested in courses that ranked higher in the recommendation list, especially the top 3 ranked courses, which 70 ~ 83% of students are interested in. However, individual differences are also found in the course selection preference of students, with some students only interested in 1 to 2 courses in the recommendation list, while most students are interested in 3 to 5 courses in the list. When students are interested in the course being recommended to them, 90% of them are based on intrinsic motivation reasons, including personal interest or attracted by the syllabi, indicating that most students approved the courses recommended by PHCRS. In contrast, 10% of them select the recommended courses based on extrinsic motivation reasons, including obtaining the necessary credits for graduation or is easier to pass the course. This study further utilizes the actual course selection data of the students to discuss whether they select courses based on the recommendation list. Moreover, the results show a considerable gap in matching between 20% to 100%, meaning that even though some students are interested in the course recommendation list, they may not consider taking those courses. The possible reasons for this may be personal or environmental interference, but there is no substantial difference in learning outcomes whether they follow the recommendation list or not, indicating that the recommended course provided by PHCRS are not necessarily those that are easier to receive good grades.

7. Contributions, limitations, and future work

The PHCRS proposed in this study proves its ability of recommending adequate course lists for students from different departments while taking human factors into consideration and providing recommendations that suit the students' needs. Only few research studies proved the effectiveness of the AI course recommendation system on students' learning outcomes (e.g., Esteban et al., 2020), but these systems focused on specific subjects by collecting additional data for their experiments. The PHCRS proposed in this study eliminates this downside by developing the system directly utilizing the course selection data, then uses AI to find out the potentials and disadvantages of students and recommend adequate courses for students to select. The system is now available for all students in NYCU. Moreover, the PHCRS database can track the learning progress and learning outcomes of students through its own database or concatenate the data from the university database and then provide recommendations by taking these data into consideration. In the future, we can adjust or expand the functionality of the PHCRS through historic data and provide interdisciplinary course recommendations and real-time learning outcome feedback, making the recommendation results more focused on the need of students in different learning stages.

Second, our research proved the potential of the genetic algorithm in finding the optimum weights of the parameters in a recommendation system, especially in chromosome modeling, in which the genetic algorithm can optimize the relative parameters, such as the size of neighbors and similarity metrics. This method can set up the best parameter setting combinations for each student. However, the recommended courses can be affected by personal preferences, course selection regulations, or the environment that the student is in, making the recommendation not 100% accurate (Chang et al., 2022; Esteban et al., 2020). This is a common restriction in human-centered recommendation systems; no state-of-the-art systems can include all algorithms, and no state-of-the-art algorithms can be applied without sacrificing accuracy in some fields (Lee et al., 2023). Although it is a tough task, to make the recommendation more accurate, we will keep using AI techniques to find out the factors affecting students' course selection decisions and their needs. By taking these human or environment factors into the construction of the recommendation model, the recommendation results can be closer to the true personal needs of students and can be more accurate.

Finally, the case study only tracks one semester of use of PHCRS, and the results indicated that those who selected courses based on the recommendations provided by PHCRS did not have higher motivations nor higher grades than those who did not. Based on the records collected by PHCRS, even though this study practiced the value of human-centered AI while developing PHCRS, there are still some issues that can be solved by further studies or system development. With the PHCRS being open to all undergraduate students, what kind of characteristics or student needs made them more intrigued to use PHCRS? Do departments with more students and courses hold higher standards towards PHCRS? Do students change their course selection preferences after using PHCRS for some time? How does PHCRS change its recommendation algorithm accordingly? Future research can concatenate with other databases of interest, adding real-time feedback or learning analysis, offering this information to students and teachers to achieve the goal of learning outcome optimization.

Acknowledgement

The authors would like to thank the National Science and Technology Council of Republic of China for financially supporting this research under Contract No. NSC 110-2222-E-008-008-MY3 and NSC 111-2410-H-A49-030-.

Reference

- Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Bendaddou, D., Ayyash, M., & Al-Fuquaha, A. (2020). *Artificial Intelligence in Education: A Panoramic review*. <https://doi.org/10.35542/osf.io/zvu2n>
- Alkhatlan, A., & Kalita, J. (2019). Intelligent tutoring systems: A Comprehensive historical survey with recent developments. *International Journal of Computer Applications*, 181(43), 1-20. <https://doi.org/10.5120/ijca2019918451>
- Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A Systematic literature review. *Intelligent Data Analysis*, 21(6), 1487-1524. <https://arxiv.org/abs/1901.03888>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7, 1247-1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chang, H. T., & Chen, H. H. (2021). The Learning outcomes in short-term advanced technology interdisciplinary hands-on courses using multigroup SEM. *Chinese Journal of Science Education*, 29(3), 245-266. [https://doi.org/10.6173/CJSE.202109_29\(3\).0003](https://doi.org/10.6173/CJSE.202109_29(3).0003)
- Chang, P. C., Lin, C. H., & Chen, M. H. (2016). A Hybrid course recommendation system by integrating collaborative filtering and artificial immune systems. *Algorithms*, 9(3), 47. <https://doi.org/10.3390/a9030047>
- Chang, H. T., Lin, C. Y., Wang, L. C., & Tseng, F. C. (2022). How students can effectively choose the right courses: Building a recommendation system to assist students in choosing courses adaptively. *Educational Technology & Society*, 25(1), 61-74.
- Defiebre, D., Sacharidis, D., & Germanakos, P. (2022). A Human-centered decentralized architecture and recommendation engine in SIoT. *User Modeling and User-Adapted Interaction*, 32(3), 297-353. <https://doi.org/10.1007/s11257-022-09320-3>
- Dwivedi, S., & Roshni VS, D. K. (2017, August 3-4). Recommender system for big data in education. In *Proceedings of 2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH)*. <https://doi.org/10.1109/ELELTECH.2017.8074993>

- Esteban, A., Zafra, A., & Romero, C. (2020). Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. *Knowledge-Based Systems*, 194(22), 1-14. <https://doi.org/10.1016/j.knosys.2019.105385>
- Farzan, R., & Brusilovsky, P. (2006). Social navigation support in a course recommendation system. In *Adaptive Hypermedia and Adaptive Web-Based Systems. AH 2006. Lecture Notes in Computer Science*, 4018, 91-100. https://doi.org/10.1007/11768012_11
- Han, J. W., Jo, J. C., Ji, H. S., & Lim, H. S. (2016). A Collaborative recommender system for learning courses considering the relevance of a learner's learning skills. *Cluster Computing*, 19, 2273-2284. <https://doi.org/10.1007/s10586-016-0670-x>
- Holland, J. (1975). *Adaptation in natural and artificial systems: An Introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press.
- Huang, M., Huang, H., & Chen, M. (2007). Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems with Applications*, 33(3), 551-564. <https://doi.org/10.1016/j.eswa.2006.05.019>
- Iatrellis, O., Kameas, A., & Fitsilis, P. (2017). Academic advising systems: A Systematic literature review of empirical evidence. *Education Science*, 7(4), 90. <https://doi.org/10.3390/educsci7040090>
- Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S., & Suparta, W. (2019). A Proposed framework in an intelligent recommender system for the college student. *Journal of Physics: Conference Series*, 1402(6), 1-7. <https://doi.org/10.1088/1742-6596/1402/6/066100>
- Lee, A. V. Y., Luco, A. C., & Tan, S. C. (2023). A Human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educational Technology & Society*, 26(1), 147-159.
- Lee, Y. M., & Sun, S. H. (2010). The relationship between autonomous motivation of course-taking and learning engagement on college students. *Journal of Research in Education Sciences*, 55(1), 155-182. <https://doi.org/10.3966/2073753X2010035501006>
- Lin, J., Pu, H., Li, Y., & Lian, J. (2018). Intelligent recommendation system for course selection in smart education. *Procedia Computer Science*, 129, 449-453. <https://doi.org/10.1016/j.procs.2018.03.023>
- Liu, Z., Zhang, N., Peng, X., Liu, S., & Yang, Z. (2023). Students' social-cognitive engagement in online discussions: An Integrated analysis perspective. *Educational Technology & Society*, 26(1), 1-15.
- Ngaffo, A. N., Ayeb, W. E., & Choukair, Z. (2020). A Bayesian inference based hybrid recommender system. *IEEE Access*, 8, 101682-101701. <https://doi.org/10.1109/ACCESS.2020.2998824>
- Popenici, S. A. D. & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Learning*, 12(22). <https://doi.org/10.1186/s41039-017-0062-8>
- Renz, A., Krishnaraja, S., & Gronau, E. (2020). Demystification of artificial intelligence in education: How much AI is really in the educational technology? *International Journal of Learning Analytics and Artificial Intelligence in Education (ijAI)*, 2(1), 14-30. <https://doi.org/10.3991/ijai.v2i1.12675>
- Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11(5), 5-16.
- Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., & Oleagordia-Ruiz, I. (2021). Recommendation systems for education: Systematic review. *Electronics*, 10(14), 1-21. <https://doi.org/10.3390/electronics10141611>
- Sawarkar, N., Raghuwanshi, M. M., & Singh, K. R. (2018). Intelligent recommendation system for higher education. *International Journal on Future Revolution in Computer Science and Communication Engineering*, 4(4), 311-320.
- Schmidt, A. (2020). Interactive human-centered artificial intelligence: A Definition and research challenges. In *Proceedings of the International Conference on Advanced Visual Interfaces*, 3, 1-4. <https://doi.org/10.1145/3399715.3400873>
- Wang, Y., Wang, L., Li, Y., He, D., Chen, W., & Liu, T. Y. (2013). A Theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, 8, 6.
- Xu, W. (2019). Toward human-centered AI: A Perspective from human-computer interaction. *Interactions*, 26(4), 42-46. <https://doi.org/10.1145/3328485>
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(39). <https://doi.org/10.1186/s41239-019-0171-0>

Zhang, H. R., Min, F., He, X., & Xu, Y. Y. (2015). A Hybrid recommender system based on user-recommender interaction. *Mathematical Problems in Engineering*, 2015, 1-12. <https://doi.org/10.1155/2015/145636>

Effects of Incorporating an Expert Decision-making Mechanism into Chatbots on Students' Achievement, Enjoyment, and Anxiety

Ting-Chia Hsu^{1*}, Hsiu-Ling Huang², Gwo-Jen Hwang^{2*} and Mu-Sheng Chen¹

¹Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taiwan // ²Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan // ckhsu@ntnu.edu.tw // hsiuling427@gmail.com // gjhwang.academic@gmail.com // mushengchen946@gmail.com

*Corresponding author

ABSTRACT: In traditional instruction, teachers generally deliver the content of textbooks to students via lectures, making teaching activities lack vibrancy. Moreover, in such a one-to-many teaching mode, the teacher is usually unable to check on individual students' learning status or to provide immediate feedback to resolve their learning problems. Chatbots provide an opportunity to address this problem. However, conventional chatbots generally serve as information providers (i.e., providing relevant information by matching keywords in a conversation) rather than as decision-making advisors (i.e., using a knowledge-base with a decision-making mechanism to help users solve problems). Thus, this study proposes an expert decision-making-based chatbot to facilitate individual students' construction of knowledge during the learning process. A quasi-experiment was conducted to compare the differences in the performances and perceptions of students using the expert decision-making-based chatbot (EDM-chatbot) and the conventional chatbot (C-chatbot) in the activities of a geography course. One class of 35 students was the experimental group, using the EDM-chatbot. The other class of 35 students was the control group, using the C-chatbot. The results of the study showed that the EDM-chatbot combined with expert decision-making knowledge significantly improved students' learning achievement and learning enjoyment as well as reducing their learning anxiety, showing the value of the proposed approach.

Keywords: Artificial Intelligence in Education, Expert knowledge, Decision tree, Chatbot, Interactive learning system

1. Introduction

In recent years, several studies have reported the benefits of using ICT in traditional instruction, such as the use of multimedia to present learning content. On the other hand, scholars have found that students generally need immediate support to help them address their misconceptions or solve any problems they encounter (Weaver, 2006). However, in a traditional classroom, the teacher may be the only person who can answer students' questions. With dozens of students in a class, it is almost impossible for teachers to provide instant feedback to individual students. Therefore, it is important to encourage students to find answers themselves using information tools. With the increasing use of Artificial Intelligence (AI) technologies in education, the main research topics include intelligent tutoring systems for special education; natural language processing for language education; educational robots for AI education; educational data mining for performance prediction; discourse analysis in computer-supported collaborative learning; neural networks for teaching evaluation; affective computing for learner emotion detection; and recommender systems for personalized learning (Chen et al., 2022). Few studies have considered humanity when employing AI in education. A previous study employed human-centered AI to give students individual responses by analyzing their learning behaviors, learning environments, or strategies (Yang, 2021). Yang (2021) pointed out that AI research in education is encountering new challenges of reshaping the research trend from technology to humanity. The climate unit is one of the most complicated learning topics for students in the discipline of geography because there are numerous conditions and requirements for judging climate classification. Giving students systematic and personalized guidance when learning this topic has become crucial. Therefore human-centered AI should be designed to support the self-learning of geography.

Self-inquiry, that is, making inquiries about questions by oneself, can increase one's learning achievement and is therefore an effective strategy for students to achieve further understanding. In the field of education, chatbots serve as a learning tool where information needed for education can be stored in a database and can be retrieved or supplemented at any time by querying the bot, either orally or through text (Wollny et al., 2021). However, if each learning note in the chatbot is independent and there is no scaffolding option for students to select, they may fall into the loop of the same Q&A cycle or miss some learning notes because they never mention the decision conditions during the conversation.

In this study, the climate unit learning content was organized and constructed so that students could learn by talking to a chatbot with two different mechanisms. Students could acquire knowledge from the chatbot and then organize that knowledge. This study aimed to reduce students' learning anxiety and maintain their learning enjoyment through chatbot learning to promote better learning outcomes. Accordingly, in this study, the control group used a C-chatbot as a teaching assistant to immediately respond to their questions by referring to the database containing each learning note. The experimental group used the EDM-chatbot which incorporated expert knowledge decision making, thus applying Artificial Intelligence in Education (AIED) to achieve adaptive learning. It was expected that the students could increase their learning achievement and enjoyment, while also reducing their learning anxiety through the use of the EDM-chatbot. The research questions in this study are as follows.

- (1) Did the students using the EDM-chatbot have better learning achievement than those using the C-chatbot?
- (2) Did the students using the EDM-chatbot have lower learning anxiety than those using the C-chatbot?
- (3) Did the students using the EDM-chatbot have better learning enjoyment than those using the C-chatbot?

2. Literature review

2.1. Artificial Intelligence in Education (AIED)

AI means the ability of computers to perform tasks by simulating intelligent human behaviors (Duan et al., 2019). AI technologies have been applied in various forms in various fields, such as medical judgment precisely through image recognition via big data (Hulsen et al., 2019), or research on user interfaces that provide personalized feedback to users with voice and gesture recognition and natural language processing, the combination of voice recognition and natural language robots for business models (Okuda & Shoda, 2018), and health management (Nadarzynski et al., 2019).

AIED provides student-centered learning and uses AI to accelerate personalized learning on the one hand, providing students with personalized learning guidance or support based on their learning status, preferences, or personal characteristics (Hwang et al., 2020). Therefore, the role of the teacher changes with the help of AI and robots to provide personalized instruction, shifting to that of a supervisor or facilitator who designs and selects machines to support the students' learning, and who monitors their learning progress (Edwards et al., 2018). Therefore, innovative and productive learning activities have been designed, and better technology-enhanced learning applications have been developed to facilitate teaching, learning, or decision making; in particular, with the help of computer systems that simulate human intelligent reasoning, judgment, or prediction, AI technologies can provide personalized instruction to students (Hwang et al., 2020). For instance, a deep learning-assisted online intelligent English teaching system was proposed to help students improve the efficiency of English teaching based on their knowledge and personality acquisition (Sun et al., 2020), while online learning with social robots was used for assisting curriculum. A previous study attempted to combine the mind-mapping-guided chatbot approach to boost students' English speaking performance. This approach led to better performance than the conventional chatbot approach (Lin & Mubarak, 2021). Based on those successful applications of AIED, one of the AI techniques, supervised machine learning and decision tree, was employed in the interactive learning environment of the current study.

2.2. Chatbots

Chatbots, also known as virtual assistants, are a primitive form of AI software that can mimic human conversations and provide users with a new form of flexibility so as to achieve instant interaction (Dahiya, 2017). For instance, the emergence of chatbots, most notably Apple Siri, Microsoft Cortana, Facebook, and IBM Watson, is becoming a common trend in many fields such as medicine, the product and service industries, and education. Chatbots have a long history of being used as teaching agents in educational settings. The chatbots led to positive learning outcomes and help provide students with better learning and a better personalized learning experience (Vanichvasin, 2021). The use of chatbots in classroom tasks can have motivational effects (Fryer et al., 2017), as well as providing access to multimedia content with portability, flexibility, and immediate searching for information (Gikas & Grant, 2013). Chatbots are not limited to time and place, but can be used for supporting learning anytime and anywhere (Shah et al., 2016). Despite the maturity of chatbot technology, there is still a need to investigate how to properly add value to human practice in education through the use of chatbot technology, including the challenge of designing effective dialogues between humans and robot technology.

Due to the large number of students enrolled in the online course, students solved problems with the support of the instant feedback given by the web bot. There was a study on combining chatbots with a game learning platform to help students enter the game and perform multiple-choice tests through interactive discussions. Nenkov (2015) implemented intelligent agents on the platform IBM Bluemix using the IBM Watson technology. Chatbots have been applied in some courses such as computer science and computer networking fundamentals courses, including for Python learning (Okonkw & Ade-Ibijola, 2020). In another study, by working with a chatbot, post-secondary writers developed a thesis statement for their argumentative essay outlines, and the chatbot helped them refine their peer review feedback (Lin & Chang, 2020). A knowledge-based chatbot system was integrated into the teaching activities of a physical examination course in nursing education, using smartphones as learning devices to guide students in practicing their anatomy knowledge and analyzing the effectiveness and enjoyment of their learning (Chang et al., 2022). The impact of a teaching simulation activity using chatbots on pre-service teacher effectiveness was studied by Song et al. (2022). Accordingly, the chatbots have been used in language learning (Fryer et al., 2017), writing skills (Lin & Chang, 2020). Accordingly, the chatbot in the current study is a task-based chatbot designed to achieve learning goals by obtaining the intention and entities in the user's messages with natural language processing (NLP), adopting a free-form textual dialogue model that does not constrain the user's choices, and allows the user to interact more naturally with the robot.

2.3. Expert systems

Expert systems research has been one of the longest running and most successful areas of AI (Wagner, 2017). An expert system is a knowledge-based program that can be used to solve problems in a specific domain and provide "professional level" answers like human experts. The methodologies used in the domain can provide much help to geographers as a means of presenting geographic knowledge in a form that is accessible to many people (Fisher, 1989). Early research, based on domain knowledge provided by experienced teachers, proposed an expert system-based instructional approach to effective context-aware ubiquitous science learning (Wu et al., 2013). Using AI technologies to simulate teachers' knowledge and experience to provide individual students with personalized supports or guidance has been recognized as a potential solution (Pai et al., 2020).

A decision tree is a classification of knowledge and the relations of the concept nodes. Concepts shown as nodes and the relationships between the tree are connected with lines, like a concept map of learning material according to the classification of expert knowledge. In this study, an EDM-based chatbot was constructed based on the learner's prior knowledge measured against the results of a pre-assessment test, and a decision tree was generated based on the prior knowledge of the learner and similar former learners who had previously completed the course. The learning path was then recommended to the learner as a personalized learning tree. Decision tree classification is an important data classification technique which represents a mapping relationship between object attributes and object values. In order to employ the expert's knowledge in the application, the expert knowledge decision tree uses decision tables and decision trees to retrieve expert knowledge. The decision tables are used to confirm the completeness and correctness of the knowledge retrieval and to present the retrieved knowledge in a rule-based manner.

2.4. The current study

Effective classroom questioning is crucial for effective teaching and learning. Student questioning is an important self-regulatory strategy with multiple benefits for teaching and learning science (Van der Meij, 1994). Questioning is important for knowledge construction, discussion, self-assessment, and cognitive curiosity, and is also useful for enhancing learning achievement. For example, mutual rhetorical strategies in reading lessons were found to improve reading comprehension (Ersianawati et al., 2018). In addition, questioning strategies enhance the memory of text details in second language learning, and the comprehension of main ideas (Liu, 2021). A previous study explored the benefits of repetitive practice of short-answer questions which could enhance students' long-term memory for subsequent improvements in learning performance (Lu et al., 2021). However, it is rare to see students asking questions in conventional classes; meanwhile, teachers do not always have enough time to answer all of the students' questions in one class with the pressure of instructional progress. Therefore, this study attempted to develop an EDM-chatbot with a decision tree by using the expert system architecture and features to optimize the conversation path between the chatbot and the students, and to help students concentrate on the learning goals and focus on the interaction.

2.5. Learning enjoyment and anxiety

Learning anxiety refers to the negative emotions that students experience during the learning process; they may feel anxious at different stages of learning (Alnuzaili & Uddin, 2020), which is a common negative emotional response of learners during the learning process. Learners with higher levels of anxiety are more burdened with learning, resulting in lower learning efficiency; however, the learning process cannot be completely free of anxiety, meaning that learners with the right level of anxiety can perform better. Andrade and Williams (2009) suggested that this anxiety, called “facilitative anxiety,” can make learners work harder and pursue better performance on tasks in class.

Enjoyment of learning is an affective orientation that stems from the pleasure and happiness that learners derive from learning activities (Shumow et al., 2013). By enhancing students’ enjoyment of learning, they may develop a high level of interest in the learning goal, which will then allow them to sustain their learning and enhance their learning experience (Jack & Lin, 2018). In this study, a chatbot was used to help students learn about climate concepts. The chatbot acted as a teacher to guide students, and it was hoped that its use would enhance students’ learning enjoyment.

3. Development of the Expert Decision Making (EDM)-based chatbot

This study used IBM Watson to build a chatbot for the geographical climate unit of a science course. Climate change is a complex environmental problem that can be used to examine students’ understanding, gained through classroom communication, of climate change and its interaction. Jakobsson et al. (2009) found in a study conducted through a written test that students’ understanding of climate change was poor. They pointed out, however, that a written test does not explicitly reveal students’ knowledge. Therefore, in the present study, it was considered that students’ understanding or meaning making of complicated issues such as climate change would be better if a communicative approach was used.

Table 1 shows examples of the expert knowledge for building the ID3 decision tree (Quinlan, 1983). There are 16 classifications (i.e., $C_1, C_2 \dots C_{16}$) of weather, composed of nine constructs (i.e., elevation, cold in winter and cool in summer, latitude, rainfall, dry season, summer dry, stationary front, needle forests, snow (no rain)) which have their own different critical feature values, as shown in Table 1.

Table 1. Illustration of examples

Features									Class
Elevation[m] (A)	Cold in winter and cool in summer (B)	Latitude (L)	Rainfall [mm] (D)	Dry season (E)	Summer Dry (F)	Stationary front (G)	Needle Forests (H)	Snow no rain (I)	and
A2	Yes	L1	D2	Yes	No	No	Yes	Yes	C_1
A2	No	L1	D2	Yes	No	No	Yes	Yes	C_2
A1	Yes	L2	D1	Yes	No	No	No	No	C_3
A1	Yes	L2	D1	No	No	No	No	No	C_4
A1	Yes	L2	D2	Yes	No	Yes	No	No	C_5
A1	Yes	L2	D2	Yes	No	No	No	No	C_6
A1	Yes	L2	D4	Yes	No	No	No	No	C_7
A1	Yes	L3	D2	No	Yes	No	No	No	C_8
A1	Yes	L3	D2	Yes	No	Yes	No	No	C_9
A1	Yes	L3	D2	Yes	Yes	No	No	No	C_{10}
A1	Yes	L3	D2	Yes	No	No	No	No	C_{11}
A1	Yes	L3	D3	Yes	No	No	No	No	C_{12}
A1	Yes	L3	D4	Yes	No	No	No	No	C_{13}
A1	No	L4	D3	Yes	No	No	No	Yes	C_{14}
A1	No	L4	D3	Yes	No	No	Yes	No	C_{15}
A1	No	L4	D3	Yes	No	No	No	No	C_{16}

Entropy is used to determine the importance of the construct which is used for classification, so as to form an effective decision tree. We can calculate the gained information of each feature shown in the following based on the training data.

- Feature A. Elevation: $A1 < 3000$; $A2 \geq 3000$
- Feature B. Cold in winter and cool in summer? Yes; No
- Feature L. Latitude: $L1 = \text{None}$, $L2 < 30$, $30 \leq L3 < 60$, $L4 \geq 60$
- Feature D. Rainfall: $D1 \geq 1500$, $500 \leq D2 < 1500$, $250 \leq D3 < 500$, $D4 < 250$
- Feature E. Dry season: Yes; No
- Feature F. Summer Dry: Yes; No
- Feature G. Stationary front: Yes; No
- Feature H Needle Forests: Yes; No
- Feature I. Snow and no rain: Yes; No

Example 1:

$$\text{Gain}(S,A) = \text{Entropy}(S) - \text{Entropy}(A) = -p_{c1} \times \log_2(p_{c1}) - p_{c2} \times \log_2(p_{c2}) - p_{c3} \times \log_2(p_{c3}) \dots - p_{c16} \times \log_2(p_{c16}) - (-p_{A1} \times \log_2(p_{A1}) - p_{A2} \times \log_2(p_{A2})) = \left(-\left(\frac{1}{16}\right) \times \log_2\left(\frac{1}{16}\right) \times 16 \right) - \left(-\left(\frac{2}{16}\right) \times \log_2\left(\frac{2}{16}\right) - \left(\frac{14}{16}\right) \times \log_2\left(\frac{14}{16}\right) \right) \cong 3.46$$

To develop a decision rule for correctly classifying training examples, ID3 performs feature tests by first selecting a feature, and then using the selected feature to classify the examples into subclasses. Next, it calculates the information entropy to determine the importance of the feature based on the following Formula 1.

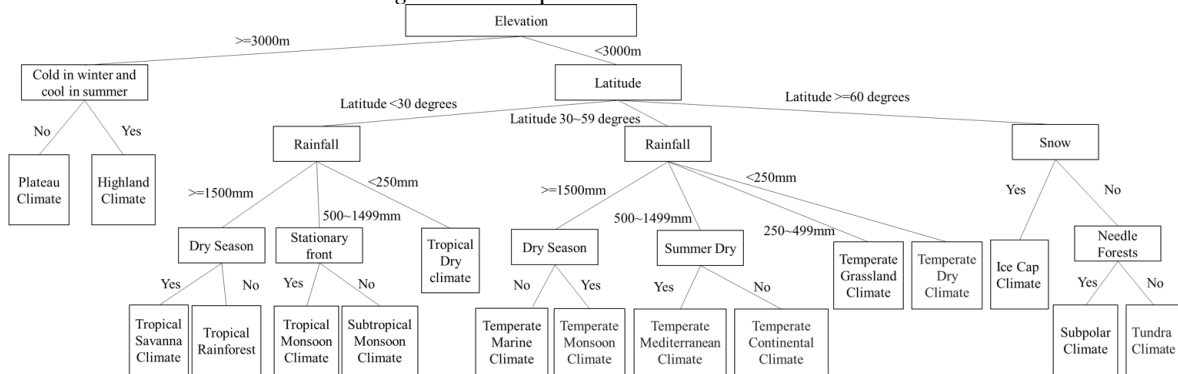
Formula 1:

$$\text{Entropy}(I) = \sum_{i=1}^C -\frac{N_i}{N} \log_2 \left[\frac{N_i}{N} \right]$$

In this formula, N , N_i , and C represent the total number of training examples, the number of examples that belong to class i , and the number of classes, respectively. Entropy can be used as an indicator of the messiness of the information quantity. The calculation of $\text{Gain}(S, A)$ indicates the profit of using attribute A (elevation in Table 1) to partition the data set S . The larger the value of Gain , the less messy the data in attribute A , and the better A can be used to classify data; the smaller the value of Gain , the greater the confusion of data in attribute A , and the worse the classification of data will be. Therefore, the information gain (S, A) represents the degree of reduction of the information complexity under the specific condition of using attribute A , equal to the information gain value of feature A . The result is calculated to be 3.46 in example 1. For example, when testing the feature “elevation,” the 16 samples are divided into two subclasses, “ ≥ 3000 ” and “ < 3000 .” Then, the sum of information entropy of each subcategory can be calculated. By subtracting the information entropy of these subclasses from the information entropy of the original training example set, ID3 deduces the information gain of the feature “elevation” as the root node at the present stage. In a similar way, the information gain for each feature can be obtained separately for testing.

When ID3 searches for features that provide the greatest information gain, the maximum information gain is obtained by comparing the gain of each feature. Next, other features are tested and the decision tree is expanded until all leaf nodes contain examples falling into a single class, as shown in Figure 1. Five vegetation groups can be distinguished as the equatorial zone, arid zone, temperate zone, cool temperate zone, and polar region. The second letter of the classification is precipitation (weather or names of climate types), and the third letter is the temperature of the location (Kottek et al., 2006).

Figure 1. Example of a climate decision tree



This chatbot has the function of learning, and adopts fuzzy matching in IBM Watson as a technique to make the conversation with students smoother. Fuzzy matching enables the system to deal with stemming, misspelling, or

partial matches. For instance, the term “running” could also be interpreted as “run,” and “bananas” could be interpreted as “banana” when dealing with the “stemming” status. Such a stemming problem occurs more in English than in Chinese. On the other hand, misspelling and partial matches more frequently occur in Chinese interactions. For example, dealing with misspelling means that even if the order of words in a phrase is incorrectly located or reversed, the original sentence can still be interpreted. “Partial match” refers to the function whereby the system is able to judge the meaning of the statement as long as certain attributes are detected in that statement. The system architecture is shown in Figure 2. The C-chatbot is shown in Figure 3. The system will search for examples and rules when it receives any questions.

Figure 2. The system architecture diagram of the EDM-chatbot

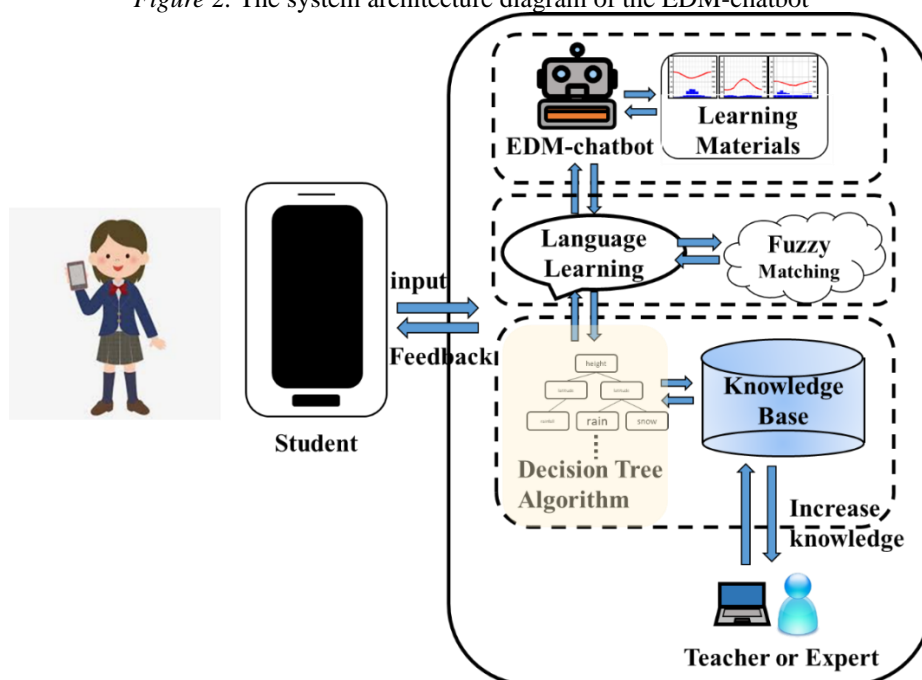
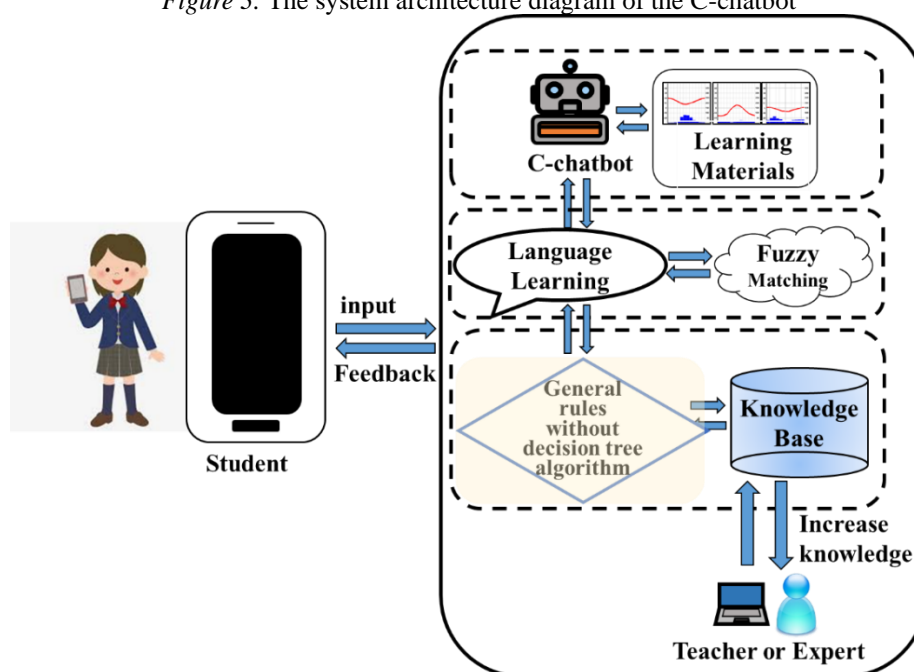


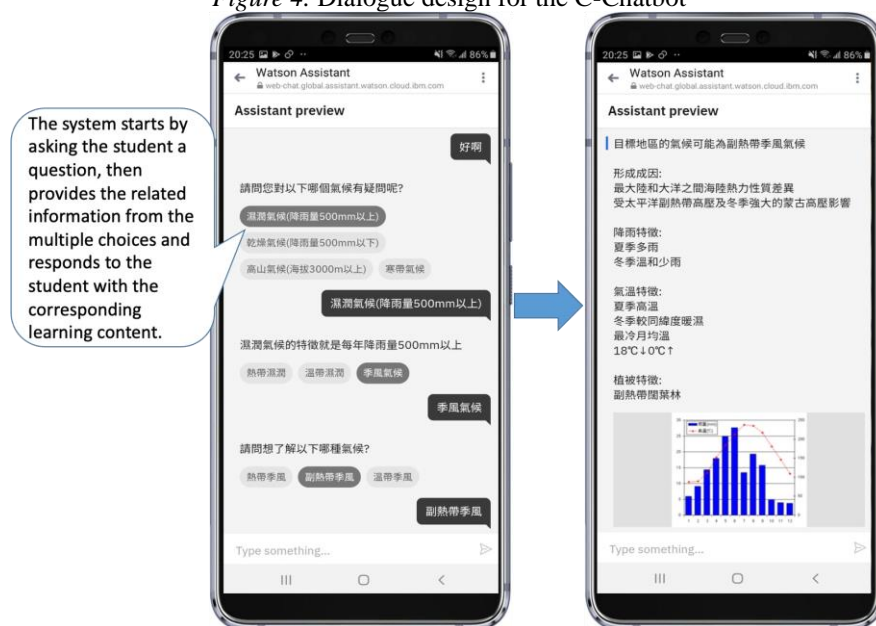
Figure 3. The system architecture diagram of the C-chatbot



The C-chatbot conversations were arranged according to the same climate feature sequences, and the dialogue replies were designed using the IBM Watson technology which can recognize similar semantics said by the students. For example, in Figure 4, the system starts by asking the student a question, then provides related information from the multiple choices, and responds to the student with the corresponding learning content in the

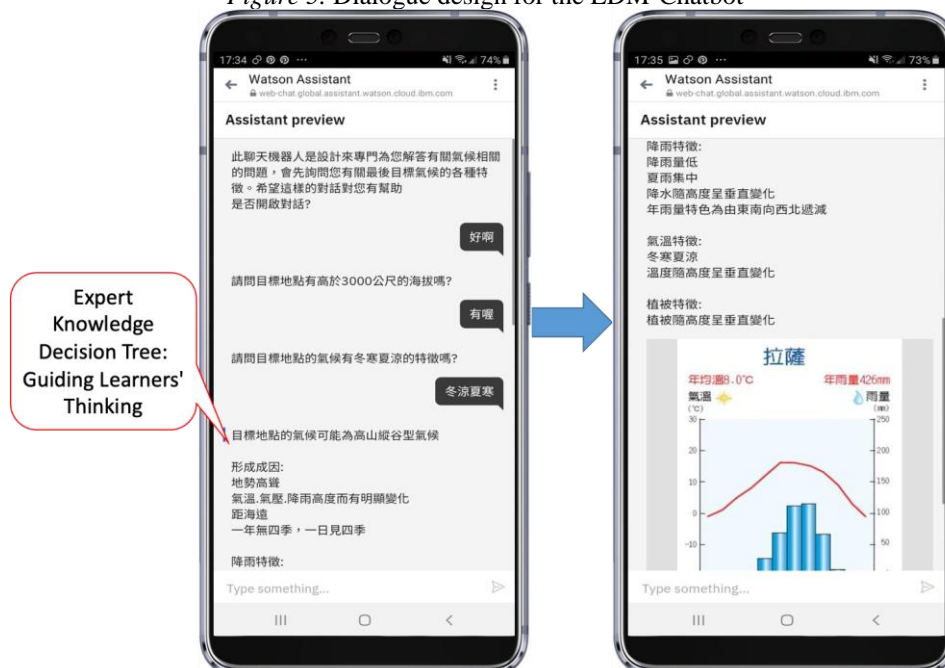
database. This is a so-called conventional chatbot. Because the C-chatbot easily falls into the same conversation loop, the example provides the conversation options for students to choose by clicking the dialogue items when they want to interact with the C-chatbot. Meanwhile, the students can also directly reply with the words they want to say if they do not just want to click the options.

Figure 4. Dialogue design for the C-Chatbot



The EDM-chatbot conversations were processed by an algorithm, so their conversations were more streamlined based on the expert knowledge and decision tree, and students were able to organize their knowledge and find their learning goals more easily. Examples comparing the two systems are shown in Figures 4 and 5.

Figure 5. Dialogue design for the EDM-Chatbot



4. Experimental design

The geographical climate expert system was designed to be used as a reference for many natural ecological studies and human activities. Each climate variable was analyzed separately for climate patterns, or data could be aggregated by using climate classifications. These classifications usually correspond to vegetation distributions,

in the sense that each climate type is dominated by a vegetation zone or an ecological region (Belda et al., 2014). Köppen was trained as a plant physiologist and believed that plants are indicators of many aspects of climate change (Belda et al., 2014). Köppen's climate classification is based on two climate elements, temperature and precipitation, and is confirmed by the distribution of natural vegetation.

4.1. Participants

In order to examine the effects of the chatbots on enhancing the learning performance of the geographical climate unit, two classes of high school students were recruited. Their average age was 17 years old. One class ($N = 35$) was the experimental group using the EDM-chatbot, while the other ($N = 35$) was the control group applying the C-chatbot. The same teacher taught both groups. The study was approved by the Research Ethics Committee of the Graduate Institute of Digital Learning and Education (approval number REA-2020-0705A). Subjects were informed that participating in the experiment was voluntary and they could withdraw from the study at any stage.

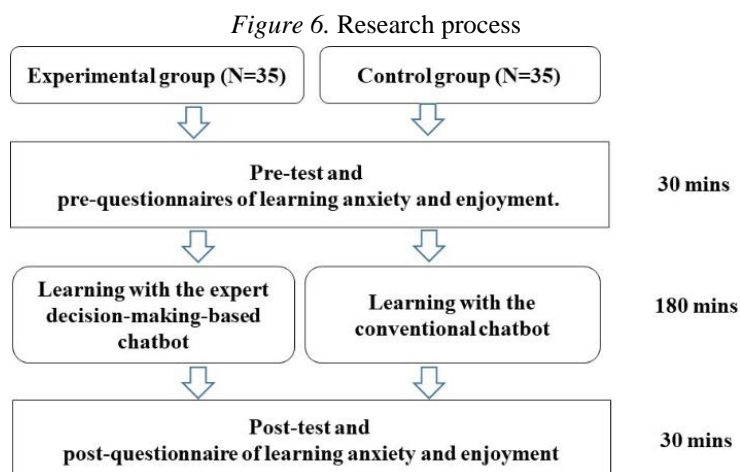
4.2. Measuring tools

For this study of applying a chatbot to the climate unit, two master students with teaching experience of 2 years on average were involved in the content development. The test of the content was jointly reviewed by two experts, and corresponded to the learning content of the chatbot. The test comprised 10 multiple-choice items, with a perfect score of 100 points in total.

The scale of learning anxiety and enjoyment used in this study was selected from the Learning Anxiety and Engagement Questionnaires (Hsu & Hwang, 2021). There are nine items in the scale of learning anxiety, which was assessed on a 5-point scale with an internal consistency reliability of 0.91. An example item is: *Learning with the chatbot makes me nervous*. There are three items in the scale of enjoyment, which was evaluated on a 5-point scale with an internal consistency reliability of 0.90. An example item is: *"The actual process of learning with the chatbot is pleasant."*

4.3. Experimental procedure

The experimental process is shown in Figure 6. Before the chatbot-based learning activity, the students took a pre-test to examine their basic knowledge related to geographical climate and filled out the learning anxiety and enjoyment questionnaires.



During the learning activity, each group spent three hours in total. The students were first guided to install the chatbot on their mobile phones and use it to complete their individual learning tasks by answering a set of questions on learning sheets prepared by the teacher. All the students in the same group used their personalized chatbot in the same classroom. Individual students needed to interact with the chatbot to get hints for the self-learning tasks during the three periods, where each period was 50 minutes with a 10-minute break between. They could talk to the chatbot via audio or text input. It should be noted that both groups were asked to complete the

same geography learning unit on climate. The only difference between the two groups was that the experimental group used the EDM-chatbot, while the control group used the C-chatbot. Both groups completed the experiment in half a day, but the two experiments were conducted on different days.

After the learning activity, the students took a post-test during which they could not use the chatbot. The learning achievement post-test comprised 10 multiple-choice items related to the knowledge of geographical climate in the chatbot. The students also completed the learning anxiety and enjoyment post-questionnaires.

After the experiment, the statistical analysis was performed. The results are presented in the next section.

5. Experimental results

The normality test was firstly carried out using the Kolmogorov-Smirnov test according to the research data; it was found that all data of each group did not conform to the normal distribution (i.e., all the p values of Shapiro-Wilk were smaller than 0.05). Therefore, the statistical methods of non-parametric analysis were conducted.

5.1. Learning achievement

First, the Wilcoxon signed-rank test was performed to compare the learning achievement pre-test and post-test of each group, as shown in Table 2. The results revealed that the learning achievement of the post-test ($M = 57.429$, $SD = 11.464$) was significantly higher than that of the pre-test ($M = 53.143$, $SD = 12.071$) in the control group ($Z = -2.044^*$, $p < .05$). Meanwhile, the learning achievement of the post-test ($M = 65.714$, $SD = 15.202$) was remarkably higher than that of the pre-test ($M = 57.143$, $SD = 23.082$) in the experimental group ($Z = -2.736^{**}$, $p < .01$). Consequently, both systems were helpful for self-learning.

Table 2. Results of the Wilcoxon signed-rank test on learning achievement for the two groups

Group	N	Pre-test		Post-test		Z
		Mean	SD	Mean	SD	
Experiment	35	57.143	23.082	65.714	15.202	-2.736**
Control	35	53.143	12.071	57.429	11.464	-2.044*

Note. ** $p < .01$; * $p < .05$.

Next, the Mann-Whitney U test was performed for comparing the pre-test of the two groups. The results confirmed that there was no significant difference between the prior knowledge of the students ($U = 485.500$; $Z = -1.527$; $p = .127 > .05$). Finally, the Mann-Whitney U test was performed again for comparing the post-test of the two groups. The results found that the learning achievement ($M = 65.714$, $SD = 15.202$) of the experimental group outperformed the learning achievement ($M = 57.429$, $SD = 11.464$) of the control group significantly ($U = 416.500$, $p < .05$), as shown in Table 3.

Table 3. Results of the Mann-Whitney U test on learning achievement for the two groups

Group	N	Mean	SD	Average Rank	Rank Sum	U	W	Z
Experiment	35	65.714	15.202	41.10	1438.50	416.500*	1046.500*	-2.364*
Control	35	57.429	11.464	29.90	1046.50			

Note. * $p < .05$.

5.2. Learning anxiety

First, the Wilcoxon signed-rank test was performed to compare the learning anxiety pre-test and post-test of each group, as shown in Table 4. The results revealed that there was no significant difference between the anxiety pre-test ($M = 3.083$, $SD = 0.439$) and the anxiety post-test ($M = 3.117$, $SD = 0.279$) of the control group ($Z = -0.432$, $p > .05$). On the contrary, there was a significant difference between the anxiety pre-test ($M = 2.844$, $SD = 0.490$) and the anxiety post-test ($M = 2.390$, $SD = 0.611$) in the experimental group ($Z = -2.893^{**}$, $p < .01$). It was found that the EDM-chatbot was helpful for significantly decreasing the students' learning anxiety.

Next, the Mann-Whitney U test was performed for comparing the learning anxiety pre-test of the two groups. The results confirmed that there was no significant difference between the prior learning anxiety of the students ($U = 454.500$; $Z = -1.883$; $p > .05$). Finally, the Mann-Whitney U test was performed again for comparing the

learning anxiety post-test of the two groups. The results found that the learning anxiety ($M = 2.390$, $SD = 0.611$) of the experimental group was lower than the learning anxiety ($M = 3.117$, $SD = 0.279$) of the control group, significantly ($U = 216.500^{***}$, $p < .001$), as shown in Table 5.

Table 4. Results of the Wilcoxon signed-rank test on learning anxiety for the two groups

Group	N	Pre-test		Post-test		Z
		Mean	SD	Mean	SD	
Experiment	35	2.844	0.490	2.390	0.611	-2.893**
Control	35	3.083	0.439	3.117	0.279	-0.432

Note. ** $p < .01$.

Table 5. Results of the Mann-Whitney U test on learning anxiety for the two groups

Group	N	Mean	SD	Average Rank	Rank Sum	U	W	Z
Experiment	35	2.390	0.611	24.19	846.50	216.500***	846.500***	-4.691***
Control	35	3.117	0.279	46.81	1638.50			

Note. *** $p < .001$.

5.3. Learning enjoyment

First, the Wilcoxon signed-rank test was performed to compare the learning enjoyment pre-test and post-test of each group, as shown in Table 6. The results revealed that the enjoyment post-test ($M = 2.790$, $SD = 0.801$) was lower than the enjoyment pre-test ($M = 3.419$, $SD = 0.711$) in the control group, significantly ($Z = -3.105^{**}$, $p < .01$). This finding revealed that the students perceived lower learning enjoyment when they carried out self-learning with the C-chatbot. On the contrary, there was no significant difference between the enjoyment pre-test ($M = 3.324$, $SD = 0.810$) and the enjoyment post-test ($M = 3.343$, $SD = 0.865$) in the experimental group ($Z = -0.082$, $p > .05$).

Table 6. Results of the Wilcoxon signed-rank test on learning enjoyment for the two groups

Group	N	Pre-test		Post-test		Z
		Mean	SD	Mean	SD	
Experiment	35	3.324	0.810	3.343	0.865	-0.082
Control	35	3.419	0.711	2.790	0.801	-3.105**

Note. ** $p < .01$.

Next, the Mann-Whitney U test was performed for comparing the learning enjoyment pre-test of the two groups. The results confirmed that there was no significant difference between the prior learning enjoyment of the students ($U = 570.000$; $Z = -0.524$; $p > .05$). Finally, the Mann-Whitney U test was performed again for comparing the learning enjoyment post-test of the two groups. The results found that the learning enjoyment ($M = 3.343$, $SD = 0.865$) of the experimental group was higher than the learning enjoyment ($M = 2.790$, $SD = 0.801$) of the control group, significantly ($U = 404.000^*$, $p < .05$), as shown in Table 7.

Table 7. Results of the Mann-Whitney U test on learning enjoyment for the two groups

Group	N	Mean	SD	Average Rank	Rank Sum	U	W	Z
Experiment	35	3.343	0.865	41.46	1451.00	404.000	1034.000	-2.566*
Control	35	2.790	0.801	29.54	1034.00			

Note. * $p < .05$.

6. Discussion

The learning discipline in the current study, geography, is one of the humanities learning subjects. This study adopted an AI chatbot as an interactive mentor for self-learning students and compared two different chatbot designs for smart phones so as to determine the contributions of expert-based decision tree chatbots with human-centered AI to the humanities learning subjects. The EDM-chatbot can provide different levels of responses from a decision tree according to students' answers. Precision education is very similar to precision medicine in that precision medicine must be tailored to each individual difference, including genes, living environment, and lifestyle (Lin et al., 2021); in the same way, each student will face different difficulties and obstacles in learning which can be addressed by precision education. Rus et al. (2013) found that the effectiveness of teaching and

learning can be improved by using an intelligent assistance system with conversational capabilities or in the form of a chatbot. The current study also proved that the chatbot used in self-learning of humanities subjects is a good means of application to promote the learning achievement of self-learning.

The C-chatbot is a passive way to perform conversation with students, although it can recognize most of the students' semantics. However, each learning note is separately stored in the database. The conversation starts from the same sequence for every student so that the students' anxiety cannot significantly decrease. They have to pay attention so as not to miss any key point or fall into the loop of the problem. The current study provided the students with the EDM-chatbot with embedded expert decisions underpinning the system so as to provide appropriate guidance for individual students and to check each learning note based on the decision tree during conversation. Thus, the application of human-centered AI could be achieved. With such a form of self-inquiry underpinned by expert decision tree scaffolding for individuals, students can systematically and actively gain relevant concepts for knowledge construction. From the perspective of meaningful learning, connecting information from different sources in an attempt to combine what they have learned is intended to reinforce meaning and enable learners to construct knowledge effectively (Dahiya, 2017). By constructing learning nodes through expert knowledge, meaningful learning is constructed, and appropriate learning paths are selected for learners to proceed in a sequential manner.

In this study, the EDM chatbot played the role of an interactive knowledge map that provided learners with learning paths, learning support for different learners, and self-adjustment. Students using the EDM chatbot to learn could make adjustments according to their needs. For example, if the student was already familiar with the classification of highland climates, he or she would then skip this classification result according to the chatting interaction and be guided to the next type of result. This is why the students showed better academic performance after self-learning with the EDM chatbot than those who used the C-chatbot, because the application of the decision tree checking during conversation became an automatic mind tool for students or scaffolding of learning nodes. In traditional education, teachers may be discriminatory in their conversations with students, even if they are unaware of it. In chatbot learning, discriminatory language is removed during the process of setting up the chatbot. If teachers pass on the wrong knowledge and do not correct it in time, it may cause learning difficulties for students. With the chatbot approach to learning, this problem can be solved by making sure that the chatbot is built to be free of knowledge errors and guidance. In sum, the EDM-chatbot group showed lower learning anxiety than the C-chatbot group because they did not need to be afraid of the level of questions they asked, and they could get the required learning responses from the robots (Babel et al., 2021). Simplifying the chatbot conversation process by means of decision trees allows students to find adaptive learning content or answers more quickly, so they will not always be in the same dialogue loop. Therefore, the EDM-chatbot can not only reduce students' learning anxiety, but can also maintain their learning enjoyment.

7. Conclusions

The core of the human-centered AIED research is to support students' learning by designing instruments which address students' learning dilemmas and provide them with equitable access to learning opportunities. In this study, an EDM-chatbot was constructed using IBM Watson, and expert decision making was incorporated into a multi-round dialogue mechanism to provide students with adaptive learning. In AI algorithmic systems, biased words related to culture, religion, and gender are avoided, providing learners with a level playing field, and new algorithms can achieve closer to human performance with intelligent analysis, diagnosis, prediction, treatment and prevention, providing adaptive learning for students (Yang, 2021). Personalizing instruction to the unique needs of learners, developing teaching strategies (Tempelaar et al., 2021), and creating human-centered learning technologies achieved the standards of precision education (Luan & Tsai, 2021). The experimental results showed that the EDM-chatbot was more effective than the C-chatbot in terms of promoting students' learning achievement, reducing their learning anxiety, and increasing their learning enjoyment. The chatbots use natural language processing to judge the focus of the students' conversation. They will not respond to students using any biased or discriminatory language, but will converse fluently and answer the climate issue first. The conversations of the chatbots in this study were centered on the learning content and were verified to contain no discriminatory language. The learning content was designed based on the textbook content and was verified by the instructor to be explanatory and reliable. Teaching requires interaction, and chatbots provide students with immediate guidance and answers, thereby increasing learning achievement and interest, and enhancing students' enjoyment of learning (Fryer et al., 2019).

Shneiderman (2020) described human-centered AI as a promising direction for designing AI systems that support human self-efficacy, promote creativity, clarify responsibility, and facilitate social participation. This

study used a chatbot to help students learn knowledge about the climate. Chatbots can solve the problems of conventional education. It is difficult for teachers to deal with the problems encountered by each student or to spend too much time on specific learning content. Students can use a chatbot to find answers on their own and to study the content they are not familiar with at any time. However, chatbots have some limitations. Chatbots are more suitable for structured or rule-based learning content. The process of building chatbots for unstructured learning content will be very complicated, and it is also difficult for students using general chatbots to organize their knowledge structure. The chatbot does not know the student's ability in advance or their learning situation during the conversation, so it may be necessary to confirm with a pop quiz, or as in this study, options to hint and guide the students' direction can be used, as in C-chatbot, or a decision tree to structure and check the learning nodes of each student can be used, as in EDM-chatbot.

Despite the positive findings, there are some limitations to the present study that should be noted. First, if the students' answers are irrelevant to the question at hand, the chatbot might have to start the conversation from the beginning, which may make the students feel impatient. In addition to system stability and accuracy adjustment, future studies are encouraged to include a machine learning mechanism to refine the chatbot's natural language processing ability by analyzing the behavioral patterns and feedback of the students using the chatbots. It would also be valuable for future research to track students' learning emotions, or to compare the difference in the effects that voice chatbots and physically human-like chatbots have on students' learning. It is recommended that future studies first collect the learning achievement and engagement of students in traditional lectures, so that the performance of the students using e-learning combined with an AI mechanism for self-learning can be compared with the performance of students taught by a teacher in a traditional lecture class which cannot take any personalized responses into consideration. Because this study compared two mechanisms under the precondition of self-learning, teachers did not intervene in students' learning in this study. Research has identified teachers' intentions to adopt AI tools in the classroom as a factor that influences the integration of AI technologies or applications into educational curriculum design (Wang et al., 2021). Therefore, teachers' perspectives on chatbots can also be explored in future studies. Future studies are encouraged to propose other research objectives and hypotheses which are different from those in this study. In other words, it is suggested that teachers become an independent variable in further studies. Another limitation of this study is that it employed chatbots in a geographical climate unit only with limited self-learning time, so it is suggested that future studies try the highly interactive design of chatbots for different disciplines and courses for a longer period of time.

Acknowledgement

This study is supported in part by the National Science and Technology Council in Taiwan under contract numbers NSTC 111-2410-H-003-168-MY3 and 111-2410-H-011-007-MY3.

References

- Alnuzailli, E. S., & Uddin, N. (2020). Dealing with anxiety in foreign language learning classroom. *Journal of Language Teaching and Research*, 11(2), 269-273. <http://dx.doi.org/10.17507/jltr.1102.15>
- Andrade, M., & Williams, K. (2009). Foreign language learning anxiety in Japanese EFL university classes: Physical, emotional, expressive, and verbal reactions. *Sophia Junior College Faculty Journal*, 29(1), 1-24.
- Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., & Baumann, M. (2021). Small talk with a robot? The Impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics*, 13, 1485-1498. <https://doi.org/10.1007/s12369-020-00730-0>
- Belda, M., Holtanová, E., Halenka, T., & Kalvová, J. (2014). Climate classification revisited: From Köppen to Trewartha. *Climate research*, 59(1), 1-13. <https://doi.org/10.3354/cr01204>
- Chang, C. Y., Kuo, S. Y., & Hwang, G. H. (2022). Chatbot-facilitated nursing education. *Educational Technology & Society*, 25(1), 15-27.
- Chen, C. H., Koong, C. S., & Liao, C. (2022). Influences of integrating dynamic assessment into a speech recognition learning design to support students' English speaking skills, learning anxiety and cognitive load. *Educational Technology & Society*, 25(1), 1-14.
- Dahiya, M. (2017). A Tool of conversation: Chatbot. *International Journal of Computer Sciences and Engineering*, 5(5), 158-161.

- Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data—evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63-71. <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>
- Edwards, C., Edwards, A., Spence, P. R., & Lin, X. (2018). I, teacher: Using artificial intelligence (AI) and social robots in communication and instruction. *Communication Education*, 67(4), 473-480. <https://doi.org/10.1080/03634523.2018.1502459>
- Ersianawati, N. L., Santosa, M. H., & Suprianti, G. (2018). Incorporating reciprocal questioning strategy and numbered heads together in reading class. *International Journal of Language and Literature*, 2(1), 19-29. <http://dx.doi.org/10.23887/ijll.v2i1.16090>
- Fisher, P. F. (1989). Expert system applications in geography. *Area*, 31(3), 279-287.
- Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An Experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, 75, 461-468. Elsevier. <https://doi.org/10.1016/j.chb.2017.05.045>
- Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior*, 93, 279-289. <https://doi.org/10.1016/j.chb.2018.12.023>
- Gikas, J., & Grant, M. M. (2013). Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media. *The Internet and Higher Education*, 19, 18-26. <https://doi.org/10.1016/j.iheduc.2013.06.002>
- Hsu, T. C., & Hwang, G. J. (2021). Interaction of visual interface and academic levels with young students' anxiety, playfulness, and enjoyment in programming for robot control. *Universal Access in the Information Society*. <https://doi.org/10.1007/s10209-021-00821-3>
- Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., & McKinney, E. F. (2019). From big data to precision medicine. *Frontiers in Medicine*, 6, 34. <https://doi.org/10.3389/fmed.2019.00034>
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 100001, 1-5. <https://doi.org/10.1016/j.caeai.2020.100001>
- Jack, B. M., & Lin, H.-S. (2018). Warning! Increases in interest without enjoyment may not be trend predictive of genuine interest in learning science. *International Journal of Educational Development*, 62, 136-147. <https://doi.org/10.1016/j.ijedudev.2018.03.005>
- Jakobsson, A., Mäkitalo, Å., & Säljö, R. (2009). Conceptions of knowledge in research on students' understanding of the greenhouse effect: methodological positions and their consequences for representations of knowing. *Science Education*, 93(6), 978-995. <https://doi.org/10.1002/sce.20341>
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3), 25-263. <https://doi.org/10.1127/0941-2948/2006/0130>
- Lin, C. J., & Mubarak, H. (2021). Learning analytics for investigating the mind map-guided AI Chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society*, 24(4), 16-35.
- Lin, H. C., Tu, Y. F., Hwang, G. J., & Huang, H. (2021). From precision education to precision medicine. *Educational Technology & Society*, 24(1), 123-137.
- Lin, M. P. C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. *Educational Technology & Society*, 23(1), 78-92.
- Liu, H. (2021). Does questioning strategy facilitate second language reading comprehension? The Effects of comprehension measures and insights from reader perception. *Journal of Research in Reading*, 44(2), 339-359. <https://doi.org/10.1111/1467-9817.12339>
- Lu, O. H., Huang, A. Y., Tsai, D. C., & Yang, S. J. (2021). Expert-authored and machine-generated short-answer questions for assessing students learning performance. *Educational Technology & Society*, 24(3), 159-173.
- Luan, H., & Tsai, C. C. (2021). A Review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1), 250-266.
- Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A Mixed-methods study. *Digital Health*, 5, 1-12. <https://doi.org/10.1177/2055207619871808>
- Nenkov, N. (2015). Implementation of a course in “Artificial Intelligence and Expert Systems” on top of a distance-learning platform. *Computer Modelling & New Technologies*, 19, 34-36.
- Novak, J. D. (1990). Concept mapping: A Useful tool for science education. *Journal of Research in Science Teaching*, 27(10), 937-949. <https://doi.org/10.1002/tea.3660271003>

- Okonkwo, C. W., & Ade-Ibijola, A. (2020). Python-bot: A Chatbot for teaching Python programming. *Engineering Letters*, 29(1), 25-34.
- Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, 54(2), 4-8.
- Pai, C. K., Liu, Y., Kang, S., & Dai, A. (2020). The Role of perceived smart tourism technology experience for tourist satisfaction, happiness and revisit intention. *Sustainability*, 12(16), 6592. <https://doi.org/10.3390/su12166592>
- Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end-games. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 463–482). Morgan Kaufmann. Springer. https://doi.org/10.1007/978-3-662-12405-5_15
- Rus, V., D'Mello, S., Hu, X., & Graesser, A. (2013). Recent advances in conversational intelligent tutoring systems. *AI Magazine*, 34(3), 42-54. <https://doi.org/10.1609/aimag.v34i3.2485>
- Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, 58, 278-295. Elsevier. <https://doi.org/10.1016/j.chb.2016.01.004>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109-124. <https://doi.org/10.17705/1thci.00131>
- Shumow, L., Schmidt, J. A., & Zaleski, D. J. (2013). Multiple perspectives on student learning, engagement, and motivation in high school biology labs. *The High School Journal*, 96(3), 232-252.
- Song, D., Oh, E. Y., & Hong, H. (2022). The Impact of teaching simulation using student chatbots with different attitudes on preservice teachers' efficacy. *Educational Technology & Society*, 25(3), 46-59.
- Sun Z., Anbarasan M., & Praveen Kumar, D. (2020). Design of online intelligent English teaching platform based on artificial intelligence techniques. *Computational Intelligence*, 37(3), 1166-1180. <https://doi.org/10.1111/coin.12351>
- Tempelaar, D., Rienties, B., & Nguyen, Q. (2021). The Contribution of dispositional learning analytics to precision education. *Educational Technology & Society*, 24(1), 109-122.
- Van Der Meij, H. (1994). Student questioning: A Componential analysis. *Learning and Individual Differences*, 6(2), 137-161. [https://doi.org/10.1016/1041-6080\(94\)90007-8](https://doi.org/10.1016/1041-6080(94)90007-8)
- Vanichvasin, P. (2021). Chatbot development as a digital learning tool to increase students' research knowledge. *International Education Studies*, 14(2), 44-53. <https://doi.org/10.5539/ies.v14n2p44>
- Wang, Y., Liu, C., & Tu, Y. F. (2021). Factors affecting the adoption of AI-Based applications in higher education: An Analysis of teachers' perspectives using structural equation modeling. *Educational Technology & Society*, 24(3), 116-130.
- Wagner, W. P. (2017). Trends in expert system development: A Longitudinal content analysis of over thirty years of expert system case studies. *Expert Systems with Applications*, 76, 85-96. <https://doi.org/10.1016/j.eswa.2017.01.028>
- Weaver, M. R. (2006). Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, 31(3), 379-394. <https://doi.org/10.1080/02602930500353061>
- Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4, 654924. <https://doi.org/10.3389/frai.2021.654924>
- Wu, P. H., Hwang, G. J., & Tsai, W. H. (2013). An Expert system-based context-aware ubiquitous learning approach for conducting science learning activities. *Educational Technology & Society*, 16(4), 217-230.
- Yang, S. J. (2021). Guest editorial: Precision education-a new challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.

Application of Artificial Intelligence Techniques in Analysis and Assessment of Digital Competence in University Courses

Tzu-Chi Yang

Institute of Education, National Yang Ming Chiao Tung University, Taiwan // tcyang.academic@gmail.com

ABSTRACT: The development of digital competence has become an important part of higher education, and digital competence assessments have attracted considerable attention and concerns. Previous studies in this area mainly focused on self-reporting and manual review methods such as questionnaires, which offer limited assessment value. To solve this issue, this study uses natural language processing (NLP)—a current promising artificial intelligence (AI) technology—to analyze syllabi for assessing digital competence in universities. Analysis results show that the proposed method can achieve an average accuracy and consistency of over 80% with excellent efficiency. Moreover, the method demonstrates high consistency with manual evaluation results ($\kappa > 0.6$) and enables automated large-scale objective assessment. In brief, the results suggest that the proposed method is efficient, effective, and reliable, making it a valuable solution for digital competence assessment. We accordingly explore the application expansion of this method in building the digital competence of universities. Furthermore, we discuss the theoretical, methodological, and applied contributions of this study.

Keywords: Digital competence, Artificial intelligence, Higher education, Text classification, Machine learning

1. Introduction

Digital applications are growing at a rapid pace and affecting people's lives, challenging the way they communicate, learn, socialize, and work. Education is an area that is most affected by this evolution, as students need to interact using digital technology (e.g., install software and work from home) in their daily life, studies, and even future careers (Olszewski & Crompton, 2020). Therefore, digital competency is important for students, and its education plays a crucial role, particularly for higher-education institutions (i.e., universities) that provide expertise in many fields. Higher education is considered a key element in digitization development (Parkes & Harris, 2002). However, there is usually a digital competence gap between university faculty and students (Chiu et al., 2021; Gonda et al., 2020). Therefore, assessing and ensuring that universities have appropriate digital competence is key to providing quality education in the present and future. Present research pertaining to digital competence in higher education is still developing and requires more attention as well as significant efforts (Müller & Mildenberger, 2021; Zhao et al., 2021).

Previous research on university digital competence assessment usually employed questionnaires and interviews as tools and showed limited results (Guo & Huang, 2021; Starkey, 2020). The limitations are due to teachers and students having different understandings of digital competence, which causes bias errors in survey results (Lucas et al., 2021). Moreover, questionnaires and interviews require considerable cooperation; consequently, implementing them regularly and continuously is difficult (Beardsley et al., 2021). Therefore, there is an urgent need for more efficient methods that ameliorate the shortcomings of the traditional assessment methods and provide more evidence of digital competence (Cabero-Almenara et al., 2021a; Weber et al., 2018). Researchers suggested that understanding how teachers integrate digital competence into teaching and curriculum content can help researchers assess digital competence (Guillén-Gámez et al., 2021). In particular, teaching methods, techniques adopted, and content taught are usually clearly described in the syllabus (Parkes & Harris, 2002). Moreover, the teaching method and course content determine the use of teaching technologies (Boss & Drabinski, 2014; Brodsky, 2017). If the syllabus describes digital competence development or requires using specific digital competence or technologies, inferring that the teacher of the course possesses the relevant digital competence and that students in the course may develop their digital competence accordingly is reasonable. Therefore, analyzing the syllabus provides objective evidence to assess the competencies that the curriculum will bring to students, including digital competencies (Boss & Drabinski, 2014; Brodsky, 2017). Syllabus analysis being an excellent solution for assessing the digital competence in universities (Çebi & Reisoğlu, 2022). However, it is a professional textual-assessment task—usually conducted manually—which is more time-consuming, labor-intensive, and difficult than questionnaire analysis (Griffith et al., 2014). Therefore, an approach to measure digital competence on a large scale is strongly needed (Hämäläinen et al., 2021).

Because of the maturity of artificial intelligence (AI) technology, it is possible to train machines to simulate human assessment methods (Ho et al., 2021) and to reinforce assessment tasks that require human expert evaluation based on textual evidence (Hong et al., 2022; Lee et al., 2023). Artificial intelligence techniques can

be developed based on human guidance to assess digital competence through explainable algorithms (e.g., text classification) that analyze specific descriptions in the syllabus. The evidence is not only reliable (Kong et al., 2023); the fairness of the results generated by AI can also help reduce the bias of different university fields. This can include the diversity of the university and serve as a bridge between educational decision-makers and experts in different fields. These AI techniques allow us to leverage the role of university education to benefit students and society (Yang et al., 2021; Gillani et al., 2023). To this end, the purpose of this study is to answer the question, “What is the effectiveness of using artificial intelligence in assessing digital competencies in university courses?” By doing so, further suggestions to researchers, educational decision-makers, and other educational stakeholders can be explored to potentially further advance HAI in this field.

2. Related works

2.1. Digital competence and higher education

Modern digital society has witnessed a dramatic change in the way people access information, communicate, and learn. Moreover, digital competence has emerged as a new term from scientific research. It can be understood as a way of using and understanding technologies and their impacts on the digital world (Becker et al., 2017) or a set of technological capabilities that effectively optimize one’s daily life (Ferrari, 2013). The European Commission defines digital competence as an ability to safely, critically, and wisely use digital technologies in work, learning, social participation, and human interactions to meet different goals (Caena & Redecker, 2019). The development of digital competence is essential for university students because they gain diverse professional knowledge. Their future work and life will inevitably involve interactions with digital technology (Burgos-Videla et al., 2021), and higher education (i.e., university) is the key to digital competence development (Olszewski & Crompton, 2020). Accordingly, considerable emphasis is placed on the prevalence and assessment of digital competencies in higher education (Spante et al., 2018; Li et al., 2021). Researchers indicated that university educators must be linked to the digital competence required by the more complex professions of the 21st century (Cabero-Almenara et al., 2021b). Moreover, instructors should integrate digital competence into their practice and professional development (Guillén-Gámez et al., 2021). Therefore, measuring the importance of digital competence in higher education has become increasingly important in educational research, particularly in curriculum design, learning activities, and teacher–student interactions (Lázaro-Cantabrana et al., 2019).

To solve the aforementioned issue, the European Commission developed DIGCOMP as a reference framework to explain the meaning of digital competence (Carretero et al., 2017). DIGCOMP defines the following areas to assess digital competence: (1) information and data literacy, (2) communication and collaboration, (3) digital-content creation (including programming), (4) safety (including digital well-being and cybersecurity related skills), and (5) problem solving (critical thinking). For example, students’ use of online discussion demonstrates communication and collaboration; completing programming projects is a typical digital-content creation competency. Owing to its validity and reliability, DIGCOMP has become the most commonly used framework for assessing digital competence in higher education (Lucas et al., 2022).

Accordingly, DIGCOMP was adopted as a framework for assessing digital competence in the present study. Moreover, most studies use questionnaires to investigate digital competencies. On the one hand, questionnaires focus on the use of specific tools, such as search engines, online bulletin boards, or systems, and are limited by the number of questionnaire items, which may not cover the full range of learning activities at universities (López-Meneses et al., 2020). On the other hand, the digital competence of all surveys is based more on the perception and self-assessment of participants than on more objective conditions (Saltos-Rivas et al., 2021). Thus, a valid and objective method to measure digital competencies in universities is currently lacking (Wang et al., 2021).

2.2. Curriculum syllabus analysis

To address the aforementioned issue, researchers indicated that a syllabus includes teaching philosophies, course content, assignments, and capabilities that can be gained by the students (Johnson, 2006; Thompson, 2007). It serves as a faculty document that defines students’ learning outcomes and the means by which they are achieved (Afros & Schryer, 2009; Habanek, 2005). Keyword comparison can provide effective analysis reports as a reference for educational decision makers (Jeffery et al., 2017). In brief, the digital competence in an educational environment reflects all learning activities related to digital competence in the learning process (Tomczyk et al.,

2020). Even if teachers or students are unaware of their own digital competence, specific descriptions in syllabi can reveal and crystallize the existence of digital competence in the curriculum (Boss & Drabinski, 2014; Hrycaj, 2017). Typical descriptions include software instruction, digital homework grading, using digital communication media, and learning systems (König et al., 2020). Moreover, in contrast to a questionnaire, which is an instantaneous response, a syllabus is provided after careful consideration by the instructor. In most cases, instructors rely on the syllabus. Hence, reviewing these documents provides objective evidence of a teacher's or student's digital competence (Lucas et al., 2022). For example, recently, an analysis of 180 course syllabi involved the investigation of teachers' digital competence and provided libraries and teachers with appropriate recommendations to assist digital competence development (Dubicki, 2019). In another analysis, a syllabus was used to determine digital competence support opportunities for teachers and develop strategic teaching promotion, showing that syllabi are a reliable way for understanding digital competence outcomes (Beuoy & Boss, 2019).

However, a comprehensive review of all courses in a school is difficult. Previous studies indicate that analyzing 1000 courses' syllabi requires at least 480 hours of team review time, not accounting for time spent on training, compiling, and analyzing data (McGowan et al., 2016). Moreover, with constantly changing syllabi, manual analysis is neither effective nor efficient. Therefore, more efficient analysis methods must be developed.

2.3. Human-centered Artificial Intelligence in Education

To address these problems, researchers have noted that there are clear distinctions in activities and their descriptions related to digital competence in the syllabus, such as utilizing software. Because of such characteristics, AI technologies (e.g., natural language processing (NLP)) can complete tasks in an accurate and efficient manner based on human recognition and domain expertise (Yang et al., 2021). In particular, AI can automatically process complex algorithms and large databases under human control. This leverages the strengths of both humans and machines, enabling them to collaborate in a way that mutually reduces blind spots and delivers high-performance applications and real creative improvements, also known as human-centered artificial intelligence (HAI) (Shneiderman, 2020). Currently, approaching AI from an educational stakeholders' (students, teachers, and leaders) perspective by considering human conditions and contexts in educational settings has gained considerable focus in HAI applications (Renz & Vladova, 2021).

Typical HAI in educational settings can be divided into several categories, including intelligent tutoring systems (e.g., personalized learning), NLP (e.g., language education and text analysis), educational robots, educational data mining (performance prediction), and affective computing (learner emotion detection) (Wang, 2021). While most HAIs in education focus on teaching and learning outcomes, researchers have noted that the manner in which education providers and institutes use AI to reinforce their functions will be an important issue in the future (Yang, 2021). NLP is considered a key area leading the AI trend because it not only mimics human understanding but also helps educational institutes and educators make interpretable and evidence-based decisions (Chang et al., 2021; Chen et al., 2022). For example, Sun and Ni (2022) used AI to analyze and identify students' text comments on an educational video resource service system, thereby significantly reducing the manual review workload. Another study by Mohammed and Omar (2020) adopted the term frequency-inverse document frequency (TF-IDF) algorithm to automatically map test questions to the appropriate bloom taxonomy cognitively and assess students' learning outcomes. Further, Yang et al. (2021a) used bidirectional encoder representations from transformers (BERT) to replace manual work to automatically assess students' text notation skills and explore the relationship with learning outcomes.

The use of AI (e.g., NLP) to facilitate syllabus analysis has been recognized as a promising approach, and there have been some research attempts recently. For example, a study by Fréchet et al. (2020) extracted various types of software used in teaching from syllabi to provide curriculum design suggestions. In another study by (Yasukawa et al., 2020), AI was used to analyze the syllabus to determine information that must be included in the syllabus and concluded that such an approach is not only credible and efficient but can also produce systematic and objective results. Accordingly, the present study uses AI to assist in syllabus analysis for assessing the digital competencies in universities.

3. Methods

The AI method used in this study involves a text-classification technique based on NLP to analyze syllabi. It includes the TF-IDF + machine learning (ML) classifier and BERT. TF-IDF + ML is the classical text-

classification method that uses word frequency as a feature to distinguish articles and is a context-independent method. Meanwhile, BERT is the most advanced text-classification technique that has been preprocessed to consider the context of words. The former has the ability to provide interpretable classification rules, while the latter can achieve excellent performance. Based on HAI perception (Riedl, 2019), both methods are used and discussed using the results herein.

3.1. Data collection and labeling

Web crawler programs were used to collect course information offered by the authors' university in the previous year. A total of 7880 syllabi (70.6% were written in Chinese and 29.4% in English) were collected. To assess digital competence, DIGCOMP 2.1—a framework proposed by the European Commission and considered a key document in assessing digital competence—was used. It has been adopted by many countries and researchers (Hernández-Martín et al., 2021). Noting that the activities in a university may not completely reflect on the DIGCOMP framework, we focused on identifying the five areas of digital competence as suggested by previous studies (López-Meneses et al., 2020; Mattar et al., 2022) rather than examining subitems in each area.

Table 1. Examples of labelled syllabi

Dc area	Course title	Syllabus digest
NA	Music and Other: On Arts and Differences	This music appreciation course explores music and the issue of differences, better known as <i>Other</i> in social science and cultural studies. In music, portraying something foreign (or <i>Other</i>) involves various complex aesthetic and technical concerns....
Area 1	Social Media and Communication Research	Social media have been deeply integrated into the lives of millions of people for a wide variety of purposes. ... In particular, in this course, you will learn important concepts, terms, and theories related to social media; <i>explore different social media sites; critically analyze possible social, political, and psychological impacts of social media use; and come up with ideas to....</i>
Area 2	Digital Technology and Language Learning	This course aims to explore various types of popular and/or cutting-edge digital technologies and their application and influence in a second and foreign language (L2/FL) teaching and learning. By the end of this course, you will be able to do the following: <i>name the most commonly used and cutting-edge technologies for L2/FL teaching and learning, elaborate the fundamental principle of implementing technologies for L2/FL teaching and learning, demonstrate how to use selected digital technologies L2/FL teaching and learning,...</i>
Area 3	Data Structures and Object-oriented Programming	There are three major themes in this course: <i>1) Understand object-oriented programming, 2) implement C++ programs to solve problems, and 3) learn and use Standard Template Library. After completing this course, you should learn the following skills: 1) design a system using classes based on system specifications...</i>
Area 4	Network Attacks & Defenses	The popularity of the computer and Internet has a rapid and enormous impact on the life of human beings. Therefore, understanding how the network functions and help improve the security and efficiency of communication is important. This course introduces network security, network defense, and network management. <i>It enables students to learn about network security systems, detection and defense algorithms, and management knowledge and skills.</i>
Area 5	High-tech Facility Design	The purpose of this course is to provide.... High-tech includes (but is not limited to) the advanced technologies applied in the fields of microelectronics,... Students will gain skills needed to meet everchanging ... <i>Use the basic theories and principles to design systems for heating, ventilation, and air conditioning (HVAC), water/air treatment, noise, and vibration mitigation. ...Establish contamination control programs for constructing, operating, and maintaining high-tech facilities. Address the issues in automatically managing the emergency, safety, and security systems. Link to the information sources for further studies in nano/micro fabrication and research.</i>

The syllabi were labeled according to these five areas. If a syllabus clearly indicates that the teacher will use or the student must use one or more of these five areas in the course, it is labeled according to the corresponding highest area of digital competence (i.e., both Area1 and Area2 are labeled as Area2). However, if the syllabus does not describe any activities related to these five areas, it is labeled “NA,” meaning not incorporating digital competence. According to the labeling results based on the preceding criteria, of the 7880 courses, 479 were labeled as Area 1, 395 as Area 2, 1541 as Area 3, 78 as Area 4, and 112 as Area 5. There were 5275 files labeled as “NA.” In addition, each syllabus was labeled by an undergraduate student and two master’s students, and their overall labeling consistency was reflected by kappa = 0.86, indicating excellent consistency. Finally, a professor with information education expertise reviewed and corrected the syllabi that were marked inconsistently. Table 1 provides examples of labeled syllabi.

3.2. Pre-processing

In the feature-extraction stage and before the classification process, the datasets were preprocessed to reduce unnecessary, repetitive, irrelevant, and noisy raw data. We wrote a python program and used jieba, NLTK, and scikit-learn to process text segmentation, stop word removal, and for feature extraction. Moreover, unnecessary data such as punctuation marks, numbers, and non-Chinese or non-English characters were also removed and all words were converted into lowercase. Words such as “the,” “a,” “an,” and “in” in English, and “是,” “因為,” and “我們” in Chinese were removed. Although English words may also exist in Chinese syllabi, these are mostly specific terms or tool names (e.g., Circuit Simulator, Music Making), and the same is true for the English syllabi. Therefore, this study does not specifically address English in the Chinese syllabus or vice versa but rather separates the training of Chinese and English syllabi.

3.3. Feature extraction and classification

After preprocessing, the TF-IDF algorithm extracted features and conducted text classification. TF-IDF is a common weighting technique for information retrieval and text mining that evaluates the importance of a word to one file set or a corpus (Dalianis, 2018). The importance of a word increases with the number of times it appears in a given file but decreases with the increasing occurrence frequency in the corpus. In addition, BERT has become a popular deep-learning method in recent years. BERT first completes model pretraining with a wide range of thematic data and many data files; then it fine-tunes the pretraining model with specific data according to various situations to achieve excellent results (Devlin et al., 2019). Therefore, TF-IDF was used in conjunction with three common ML classifiers: support vector machines (SVM), k-nearest neighbors (KNN), and naive Bayes (NB). BERT served as a classification method.

3.4. Evaluation metrics

Accuracy, precision, recall, F1-score, and kappa value were used to evaluate the effectiveness of the abovementioned classification methods. Accuracy reflects the percentage of correctly classified syllabi from the total number of syllabus files and is the most basic classification evaluation index. F1-score, or the harmonic average of sensitivity and accuracy, provides another general indicator of model effectiveness. The kappa value evaluates the consistency of the classifications performed. All the abovementioned indicators are scored between zero and one, where zero indicates poor performance and one indicates good performance. To measure the proposed model’s effectiveness, 20% of the data not included in the training set were evaluated as a test set. We also divided all the data into 10 equal parts; for each group, we took it as test data and the remaining nine groups as training data. Thus, the 10-fold cross-validation method could be used to evaluate classifier effectiveness for preventing model overfitting (Kohavi, 1995).

4. Results and discussion

4.1. Effectiveness evaluation of syllabus analysis

Results show that when 20% of the data were used as the test set, the SVM, KNN, NB, and BERT classification accuracies ranged from 0.57 to 0.83, the F1-scores ranged from 0.59 to 0.84, and the kappa values ranged from 0.20 to 0.64 (Table 2). When the TF-IDF and ML methods were used, the SVM, F1-score, and kappa value were the highest with the test dataset or 10-fold cross-validation. Therefore, SVM exhibited the best performance in

syllabus analysis and BERT had the best overall classification effectiveness. When 10-fold cross-validation was used, the accuracy of the TF-IDF + ML models ranged from 0.68 to 0.83, which was slightly greater than that of the F1-score and kappa value. Similarly, SVM afforded the highest accuracy (0.83) among different ML methods. Although KNN was slightly less accurate than SVM, it still showed good consistency (0.59). This result shows that there was no significant difference between the two ML models. The F1-score indicates that the TF-IDF + SVM models can achieve good performance. Further, the TF-IDF + NB models performed poorly among ML models. This finding is consistent with past results because the stability of NB effectiveness is often used as the basis for text classification, and there was no outstanding effectiveness in the text classification (Xu, 2018). Nevertheless, we suggest that NB can be used as the basis for model comparison.

Table 2. Evaluation of classification models

		20% test set validation					10-fold cross-validation				
		ACC	Precision	Recall	F1	Kappa	ACC	Precision	Recall	F1	Kappa
TF-IDF+	SVM	0.65	0.63	0.65	0.64	0.47	0.83	0.55	0.60	0.57	0.50
	KNN	0.57	0.59	0.55	0.59	0.20	0.80	0.42	0.50	0.46	0.59
	NB	0.59	0.65	0.61	0.62	0.32	0.68	0.51	0.53	0.52	0.43
BERT		0.83	0.83	0.86	0.84	0.64	0.80	0.61	0.74	0.67	0.56

Table 3. Examples of syllabus files classified by AI with different digital competence areas

Dc area	Course title	Syllabus digest
NA	Experiments in Physical Chemistry	Implement the physical chemistry experiment course. Teach undergraduate students basic concepts and theories of physical chemistry. Help students understand experimental methods and skills to validate theories and experiments. Help students further understand experimental processes and principles.
Area 1	Anthropology	Introduce course description, course material use, academic performance evaluation, cultural anthropology <i>online resources</i> , and other anthropology library resources.
Area 2	Media Psychology	Media technologies are inextricably intertwined with everyone's life. They affect the ways people learn, think, interact with others, feel, and act. Understand contemporary media use, its underlying causes and mechanisms, and possible impacts. <i>Guide students to observe and think about the relationships among media technologies, people, and the social environment with mutual impacts.</i> Mid-term and final assessment reports by teams are required.
Area 3	Introduction to Computers and Programming	Fundamentals are introduced. The objective is to <i>enable students to possess the following capabilities: (1) understanding concepts and skills of C programming and (2) proficiency in solving computing tasks by programming.</i>
Area 4	Enterprise Cybersecurity	This course explores current security challenges in enterprise operation and analyzes new generations of corporate security measures, including (1) <i>status of security threats</i> , (2) <i>forward-looking defensive strategies</i> , (3) <i>security maturity assessment and defensive strategies</i> , and (4) building a strong security-management team. Case studies are included. Capital security risk assessment criteria are briefly introduced. <i>Automated tool usage is introduced to facilitate hands-on practice for students.</i>
Area 5	Computer Networks	This course introduces innovation and application capabilities of information technologies and mathematics knowledge. The following knowledge and capabilities are taught/trained: information technology tools' applications; <i>design and evaluation of computerized systems, programs, and components; identifying, analyzing, and solving problems; learning current issues; understanding the impacts of information technologies on the environment, society, and world; continuous learning; understanding professional ethics and social responsibility.</i>

Moreover, compared with the TF-IDF + ML method, the BERT model achieved the highest efficiency in almost all the criteria when the 20% nonrepeating test dataset and 10-fold cross-validation were used. The BERT model's accuracy in the 10-fold cross-validation was slightly less than that of SVM, suggesting the superiority of BERT to traditional ML classifiers in syllabus classification. In particular, the BERT accuracy was 0.83 for nonrepeating datasets, almost 1.4 times the ML model accuracy. The consistency of the BERT model was 0.64, 1.74 times higher than the best ML model (SVM). These results highlight BERT's excellent capability in syllabus analysis. Unsurprisingly, as BERT has pretrained universal language models using a cross-domain text corpus, BookCorpus, and Wikipedia, it demonstrates excellent performance in NLP tasks (Yu et al., 2019).

However, there is no large difference in the performance of classical (TF-IDF + ML) and advanced (BERT) NLP methods in classifying syllabi. A possible reason is that while TF-IDF extracts features from word frequencies, the terms/words associated with digital competence are often unique and—to a certain extent—reflect the digital competence area to which a syllabus relates. Thus, although TF-IDF does not consider the context, it still performs well compared to BERT. For example, “Python” has two different meanings: a programming language or a snake genus, but if a syllabus mentions both “Python” and “syntax” we can obviously identify that it is related to digital competence (programming language). We also found that such a finding is revealed when classifying articles in many subject domains (Kim et al., 2022).

In short, the above discussion indicates that using AI (i.e., TF-IDF + ML, BERT) to analyze syllabi can provide an average accuracy of over 80% and a consistency score greater than 0.6, which are satisfactory. Moreover, after the four models used in this study were trained, the longest time to perform a classification task was only seven minutes (using Google Colab Pro, GPU: Tesla P100, Memory: 16 GB). By contrast, manual analysis takes from a few days to more than a week. Therefore, AI methods can achieve good results similar to those of manual analysis in considerably less time and with acceptable consistency, demonstrating efficient and effective syllabus analysis capability. Table 3 lists the syllabus files classified by AI. Each classified syllabus file has a clear description corresponding to labeled digital competence areas. Nonetheless, one should be aware of the possible implications and treatment of imbalanced data, for example, by involving experts to determine which of these rare instances may be the most efficient solution to the current categorical imbalance classification model (Haixiang et al., 2017).

4.2. Digital competence assessment in universities

According to previous literature, AI methods can provide accurate, consistent, and verifiable assessment for educational data analysis (Guan et al., 2020). The presented results confirm this point and allow researchers to further explore the utilization of AI methods to assess digital competence in universities. In this study, digital competence levels of different courses were compared, e.g., differences between school levels (undergraduate/graduate schools) and among different colleges within a university. Table 4 reveals that 34% of the courses assessed contain some degree of digital competence. This is not considered low and is reasonable because the university is known for electrical engineering, electronics, and information technologies, which inevitably require digital tools.

The undergraduate courses offering digital competence are classified as Area 1, 2, and 3. The percentage of graduate courses offering digital competence is higher than that at the undergraduate level, and 25% of the courses are categorized as Area 3 (digital-content creation). This aligns with the university’s graduate school training that emphasizes independent thinking with innovative ideas. More than 80% of the courses offered by the College of Intelligent Sciences and Green Energy and more than 85% of the courses offered by the College of Information Technology require use of digital competence. By contrast, less than 20% of the College of Science courses and less than 12% of the College of Dentistry courses require use of digital competence. The College of Information Technology naturally requires extensive use of digital competence, which was clearly stated in the syllabi. By contrast, digital competence is not so widely applied in the medical and health fields, explaining the lack of digital competence displayed by the College of Dentistry and confirming results from previous studies (Golz et al., 2021; Lázaro-Cantabrana et al., 2019).

The results of this study demonstrated an HAI application that universities can use this approach to periodically review the status of digital competencies on campus. By doing so, in addition to providing evidence beyond the questionnaire response, further identification of programs for improving the digital competencies of faculty, staff, and students based on objective evidence (i.e., syllabus) is possible. This result also indicates that different universities are often organized with similar domains of expertise and provide the same courses (e.g., Microelectromechanical Systems and calculus), and that the syllabi of these courses usually have common specific terms. Accordingly, the approach adopted in this study reveals an opportunity for other higher-education institutions to demonstrate generalizability.

Although there has been some research on syllabus analysis, the expert-based approach is limited by human resources and the technique-based approach may lack involved domain knowledge. This study uses both classical (i.e., TF-IDF + ML) and advanced NLP techniques (i.e., BERT) to complete the same task. The former may provide easy-to-interpret classification rules based on word frequencies, while the latter can provide higher accuracy through repeated validation, both providing significant improvements in efficiency and consistency. This means that human experts can themselves decide the level of AI intervention to maximize their own capabilities (Shneiderman, 2020) either by leaving the analysis of the syllabus entirely to the machine or by

determining the level of automation to provide explorable results or evidence for educational decisions (e.g., looking at the proportion of digital competence and essential learning/teaching activities in each domain offered by different colleges). To the best of our knowledge, this study is the first to use HAI to assess digital competencies, which adds to the usefulness and value of HAI in the educational domain.

Table 4. Courses with digital competence in the entire university at different school levels and in different colleges

	Area1		Area2		Area3		Area4		Area5		NA	
	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>
<i>Campus</i>	6.08%	479	5.01%	395	19.56%	1541	0.99%	78	1.42%	112	66.94%	5275
<i>School</i>												
Graduate	3.30%	137	4.39%	182	25.14%	1043	1.21%	50	1.21%	50	64.76%	2687
Undergraduate	9.20%	342	5.73%	213	13.31%	495	0.75%	28	1.67%	62	69.35%	2579
<i>College</i>												
Humanities and Social Science	6.25%	30	11.88%	57	13.13%	63	0.42%	2	1.04%	5	67.29%	323
Engineering	4.09%	28	5.55%	38	13.87%	12	0.00%		1.46%	10	75.04%	514
Dentistry	1.69%	3	2.81%	5	6.74%	29	0.00%		0.00%		88.76%	158
College of Life Sciences	2.24%	8	3.08%	11	8.12%	103	0.00%		0.00%		86.55%	309
Biological Science and Technology	4.42%	24	2.95%	16	18.97%	104	0.18%	1	0.00%		73.48%	399
Biomedical Science and Engineering	3.62%	23	1.57%	10	16.35%	32	0.16%	1	0.16%	1	78.14%	497
Photonics	5.76%	11	7.85%	15	16.75%	51	0.00%		1.05%	2	68.59%	131
Industry Academic Innovation School	9.31%	39	0.48%	2	12.17%	15	0.00%		1.91%	8	76.13%	319
Hakka Studies	12.18%	24	9.14%	18	7.61%	2	1.52%	3	3.55%	7	65.99%	130
Law	9.30%	4	4.65%	2	4.65%	3	0.00%		0.00%		81.40%	35
Semiconductor Technology	2.50%	1	7.50%	3	7.50%	38	0.00%		0.00%		82.50%	33
Sciences	5.89%	31	5.32%	28	7.22%	80	0.57%	3	0.00%		80.99%	426
Artificial Intelligence	0.00%		0.00%		74.77%	420	2.80%	3	0.00%		22.43%	24
Computer Science	2.59%	16	2.10%	13	67.96%	208	6.96%	43	6.31%	39	14.08%	87
Electrical and Computer Engineering	16.77%	139	10.98%	91	25.09%	123	0.97%	8	4.22%	35	41.98%	348
Management	8.06%	66	7.33%	60	15.02%	102	1.59%	13	0.49%	4	67.52%	553
Medicine	2.70%	21	1.80%	14	13.13%	8	0.13%	1	0.13%	1	82.11%	638
Pharmaceutical Sciences	0.00%		1.16%	2	4.65%	45	0.00%		0.00%		94.19%	162
Nursing	3.65%	7	3.13%	6	23.44%	8	0.00%		0.00%		69.79%	134
Other	5.63%	4	4.23%	3	11.27%		0.00%		0.00%		77.46%	55

5. Conclusions

Assessment of digital competency in higher education is still a nascent topic. To address the limitations of the use of self-reporting and the inefficiencies of manual analysis. This study explored the following question: “What is the effectiveness of using artificial intelligence in assessing digital competencies in university courses?” from an HAI perspective. Our results point to a high degree of consistency in human analyses conducted using AI. Our results show that universities can use this approach to proactively and efficiently assess all university courses with minimal human effort. There will be an opportunity to provide equitable digital competency education to students from diverse backgrounds, resulting in greater benefits for individuals, educational institutions, and society. Based on the result, we summarized the findings and contributions of this study from three perspectives:

Regarding theory, from an educational research perspective, a syllabus represents the contract between teacher and student and reflects the activities that occur in the curriculum, and it can be an objective method of assessing specific competencies. This study uses HAI to practicalize this perspective. To the best of our knowledge, this is the first study to adopt HAI to assess digital competencies through syllabus analysis, which may provide inspiration for practicing HAI in the education field. Regarding methods, we used both classical (TF-IDF) and advanced (BERT) AI (i.e., NLP) techniques, showing that advanced AI achieves higher accuracy rates, but the classical one may provide interpretable results with acceptable accuracies. Both classical and advanced AIs significantly reduce the task time and produce reliable results. Therefore, educators can decide which AI technique to use and achieve their goals. As mentioned by Shneiderman (2020) HAI retains manual control where appropriate, thereby increasing performance and enabling creative improvements. Regarding application, this study provides an opportunity to fill the diversity and inclusion gap by establishing a joint dialogue on digital competency education among departments of different professional backgrounds in the university from the HAI perspective. We show that this approach is explainable and trustworthy in universities, and it can proactively and efficiently evaluate programs across the university with a minimal workload. Such an approach may help universities provide equitable digital competency education to students from different backgrounds, creating greater benefits and societal interests in higher education (Yang et al., 2021b). Universities will also have more opportunities to promote quality education, as emphasized in the Sustainable Development Goals (SDGs).

6. Limitations and future works

Although the results of this study are promising, the proposed has method limitations. First, this study focused on the syllabus' textual description, but the contextual relevance, semantics, and implied intention between sentences were not considered in the model. Future research could improve the performance of the classifier using other algorithms. In addition, this study assesses digital competence using a syllabus, and verifying the consistency of this method with student/instructor's perceptions of digital competence and its applicability to other universities as well as exploring the existence of overfitting effects in future research is useful. Second, this study presents a method to investigate the digital competencies in universities, although it can be used to identify solutions that facilitate the development of digital competencies for universities. However, the development of teachers' and students' digital competence may be related to individual differences such as age and gender (Gnambs, 2021). We need to clarify these relationships in future research to create effective digital-competency training programs. Finally, development of machines to understand human socio-cultural norms and theories of the mind is in its nascency, and we agree that AI cannot replace humans but rather reinforces human capabilities. Thus, this study does not address some problems (unbalanced data) but leaves the final judgment to experts to accommodate the two-dimensional framework of HAI (Shneiderman, 2020).

Acknowledgement

This research was supported by the Higher Education Sprout Project of National Yang Ming Chiao Tung University (NYCU) and the Ministry of Education (MOE), Taiwan, as well as the Ministry of Science and Technology in Taiwan through Grant numbers MOST 108-2511-H-009-019-MY2 and 111-2410-H-A49-029. We would also thank Yu Chen Wu for her support for the study.

References

- Afros, E., & Schryer, C. F. (2009). The Genre of syllabus in higher education. *Journal of English for Academic Purposes*, 8(3), 224–233. <https://doi.org/10.1016/j.jeap.2009.01.004>
- Beardsley, M., Albó, L., Aragón, P., & Hernández-Leo, D. (2021). Emergency education effects on teacher abilities and motivation to use digital technologies. *British Journal of Educational Technology*, 52(4), 1455–1477. <https://doi.org/10.1111/bjet.13101>
- Becker, S. A., Cummins, M., Davis, A., Freeman, A., Hall, C. G., & Ananthanarayanan, V. (2017). NMC Horizon Report: 2017 Higher Education Edition. *The New Media Consortium*. <https://www.learntechlib.org/p/174879/>
- Beuoy, M., & Boss, K. (2019). Revealing instruction opportunities: A Framework-based rubric for syllabus analysis. *Reference Services Review*, 47(2), 151–168. <https://doi.org/10.1108/RSR-11-2018-0072>
- Boss, K., & Drabinski, E. (2014a). Evidence-based instruction integration: A Syllabus analysis project. *Reference Services Review*, 42(2), 263–276. <https://doi.org/10.1108/RSR-07-2013-0038>

- Brodsky, M. (2017). Understanding data literacy requirements for assignments: A Business school syllabus study. *International Journal of Librarianship*, 2(1), 3–15. <https://doi.org/10.23974/ijol.2017.vol2.1.25>
- Burgos-Videla, C. G., Castillo Rojas, W. A., López Meneses, E., & Martínez, J. (2021). Digital competence analysis of university students using latent classes. *Education Sciences*, 11(8), 385. <https://doi.org/10.3390/educsci11080385>
- Cabero-Almenara, J., Barroso-Osuna, J., Gutiérrez-Castillo, J.-J., & Palacios-Rodríguez, A. (2021a). The Teaching digital competence of health sciences teachers. a study at Andalusian Universities (Spain). *International Journal of Environmental Research and Public Health*, 18(5), 2552. <https://doi.org/10.3390/ijerph18052552>
- Cabero-Almenara, J., Guillén-Gámez, F. D., Ruiz-Palmero, J., & Palacios-Rodríguez, A. (2021b). Digital competence of higher education professor according to DigCompEdu. Statistical research methods with ANOVA between fields of knowledge in different age ranges. *Education and Information Technologies*, 26(4), 4691–4708. <https://doi.org/10.1007/s10639-021-10476-5>
- Caena, F., & Redecker, C. (2019). Aligning teacher competence frameworks to 21st century challenges: The Case for the European Digital Competence Framework for Educators (Digcompedu). *European Journal of Education*, 54(3), 356–369. <https://doi.org/10.1111/ejed.12345>
- Carretero, S., Vuorikari, R., & Punie, Y. (2017). *DigComp 2.1. The Digital competence framework for citizens. With eight proficiency levels and examples of use*. Publications Office of the European Union.
- Çebi, A., & Reisoğlu, İ. (2022). Adaptation of self-assessment instrument for educators' digital competence into Turkish culture: A Study on reliability and validity. technology, knowledge and learning. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-021-09589-0>
- Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education. *Educational Technology & Society*, 25(1), 28–47.
- Chiu, T.-F., Chu, D., Huang, S.-J., Chang, M., Liu, Y., & Lee, J. J. (2021). Facing the coronavirus pandemic: An Integrated continuing education program in Taiwan. *International Journal of Environmental Research and Public Health*, 18(5), 2417. <https://doi.org/10.3390/ijerph18052417>
- Dalianis, H. (2018). Computational methods for text analysis and text classification. In H. Dalianis (Ed.), *Clinical Text Mining: Secondary Use of Electronic Patient Records* (pp. 83–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-78503-5_8
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. PsyArXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Dubicki, E. (2019). Mapping curriculum learning outcomes to ACRL's Framework threshold concepts: A Syllabus study. *The Journal of Academic Librarianship*, 45(3), 288–298. <https://doi.org/10.1016/j.acalib.2019.04.003>
- Ferrari, A. (2013). *DIGCOMP: A Framework for developing and understanding digital competence in Europe*. Publications Office of the European Union Luxembourg.
- Fréchet, N., Savoie, J., & Dufresne, Y. (2020). Analysis of text-analysis syllabi: Building a text-analysis syllabus using scaling. *PS: Political Science & Politics*, 53(2), 338–343.
- Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2023). Unpacking the “Black Box” of AI in Education. *Educational Technology & Society*, 26(1), 99–111.
- Gnams, T. (2021). The Development of gender differences in information and communication technology (ICT) literacy in middle adolescence. *Computers in Human Behavior*, 114, 106533. <https://doi.org/10.1016/j.chb.2020.106533>
- Golz, C., Peter, K. A., Müller, T. J., Mutschler, J., Zwakhalen, S. M. G., & Hahn, S. (2021). Technostress and digital competence among health professionals in Swiss psychiatric hospitals: Cross-sectional study. *JMIR Mental Health*, 8(11), e31408. <https://doi.org/10.2196/31408>
- Gonda, D., Ďuriš, V., Pavlovičová, G., & Tirpáková, A. (2020). Analysis of factors influencing students' access to mathematics education in the form of MOOC. *Mathematics*, 8(8), 1229. <https://doi.org/10.3390/math8081229>
- Griffith, S. M., Domenech Rodríguez, M. M., & Anderson, A. J. (2014). Graduate ethics education: A Content analysis of syllabi. *Training and Education in Professional Psychology*, 8(4), 248–252. <https://doi.org/10.1037/tep0000036>
- Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A Twenty-year data-driven historical analysis. *International Journal of Innovation Studies*, 4(4), 134–147. <https://doi.org/10.1016/j.ijis.2020.09.001>
- Guillén-Gámez, F. D., Mayorga-Fernández, M. J., Bravo-Agapito, J., & Escribano-Ortiz, D. (2021). Analysis of teachers' pedagogical digital competence: Identification of factors predicting their acquisition. *Technology, Knowledge and Learning*, 26(3), 481–498. <https://doi.org/10.1007/s10758-019-09432-7>
- Guo, J., & Huang, J. (2021). Information literacy education during the pandemic: The Cases of academic libraries in Chinese top universities. *The Journal of Academic Librarianship*, 47(4), 102363. <https://doi.org/10.1016/j.acalib.2021.102363>

- Habaneck, D. V. (2005). An Examination of the integrity of the syllabus. *College Teaching*, 53(2), 62–64. <https://doi.org/10.3200/CTCH.53.2.62-64>
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73, 220-239.
- Hämäläinen, R., Nissinen, K., Mannonen, J., Lämsä, J., Leino, K., & Taajamo, M. (2021). Understanding teaching professionals' digital competence: What do PIAAC and TALIS reveal about technology-related skills, attitudes, and knowledge? *Computers in Human Behavior*, 117, 106672. <https://doi.org/10.1016/j.chb.2020.106672>
- Hernández-Martín, A., Martín-del-Pozo, M., & Iglesias-Rodríguez, A. (2021). Pre-adolescents' digital competences in the area of safety. Does frequency of social media use mean safer and more knowledgeable digital usage? *Education and Information Technologies*, 26(1), 1043–1067. <https://doi.org/10.1007/s10639-020-10302-4>
- Ho, I. M. K., Cheong, K. Y., & Weldon, A. (2021). Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques. *PLOS ONE*, 16(4), e0249423. <https://doi.org/10.1371/journal.pone.0249423>
- Hong, S., Kim, J., & Yang, E. (2022). Automated text classification of maintenance data of higher education buildings using text mining and machine learning techniques. *Journal of Architectural Engineering*, 28(1), 04021045. [https://doi.org/10.1061/\(ASCE\)AE.1943-5568.0000522](https://doi.org/10.1061/(ASCE)AE.1943-5568.0000522)
- Hrycaj, P. L. (2017). An Analysis of online syllabi for credit-bearing library skills courses. *College & Research Libraries*, 67(6), 525-535. <https://doi.org/10.5860/crl.67.6.525>
- Jeffery, K. M., Houk, K. M., Nielsen, J. M., & Wong-Welch, J. M. (2017). Digging in the mines: Mining course syllabi in search of the library. *Evidence Based Library and Information Practice*, 12(1), 72–84. <https://doi.org/10.18438/B8GP81>
- Johnson, C. (2006). Best practices in syllabus writing. *The Journal of Chiropractic Education*, 20(2), 139–144.
- Kim, M. G., Kim, M., Kim, J. H., & Kim, K. (2022). Fine-tuning BERT models to classify misinformation on garlic and COVID-19 on Twitter. *International Journal of Environmental Research and Public Health*, 19(9), 5126. <https://doi.org/10.3390/ijerph19095126>
- Kohavi, R. (1995). A Study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2, 1137–1143.
- Kong, S.-C., Cheung, W. M.-Y., & Zhang, G. (2023). Evaluating an artificial intelligence literacy programme for developing university students' conceptual understanding, literacy, empowerment and ethical awareness. *Educational Technology & Society*, 26(1), 16-30.
- König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: Teacher education and teacher competence effects among early career teachers in Germany. *European Journal of Teacher Education*, 43(4), 608–622. <https://doi.org/10.1080/02619768.2020.1809650>
- Lázaro-Cantabrana, J. L., Usart-Rodríguez, M., & Gisbert-Cervera, M. (2019). Assessing teacher digital competence: The Construction of an instrument for measuring the knowledge of pre-service teachers. *Journal of New Approaches in Educational Research*, 8(1), 73–78. <https://doi.org/10.7821/naer.2019.1.370>
- Lee, A. V. Y., Luco, A. C., & Tan, S. C. (2023). A Human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educational Technology & Society*, 26(1), 147-159.
- Li, Y., Chen, Y., & Wang, Q. (2021). Evolution and diffusion of information literacy topics. *Scientometrics*, 126(5), 4195–4224. <https://doi.org/10.1007/s11192-021-03925-y>
- López-Meneses, E., Sirignano, F. M., Vázquez-Cano, E., & Ramírez-Hurtado, J. M. (2020). University students' digital competence in three areas of the DigCom 2.1 model: A Comparative study at three European universities. *Australasian Journal of Educational Technology*, 36(3), 69-88.
- Lucas, M., Bem-haja, P., Santos, S., Figueiredo, H., Ferreira Dias, M., & Amorim, M. (2022). Digital proficiency: Sorting real gaps from myths among higher education students. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.13220>
- Lucas, M., Bem-Haja, P., Siddiq, F., Moreira, A., & Redecker, C. (2021). The Relation between in-service teachers' digital competence and personal and contextual factors: What matters most? *Computers & Education*, 160, 104052. <https://doi.org/10.1016/j.compedu.2020.104052>
- Mattar, J., Ramos, D. K., & Lucas, M. R. (2022). DigComp-based digital competence assessment tools: Literature review and instrument analysis. *Education and Information Technologies*, 1-25. <https://doi.org/10.1007/s10639-022-11034-3>
- McGowan, B., Gonzalez, M., & Stanny, C. J. (2016). What do undergraduate course syllabi say about information literacy? *Portal: Libraries and the Academy*, 16(3), 599–617. <https://doi.org/10.1353/pla.2016.0040>

- Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS one*, 15(3), e0230442.
- Müller, C., & Mildenberger, T. (2021). Facilitating flexible learning by replacing classroom time with an online learning environment: A Systematic review of blended learning in higher education. *Educational Research Review*, 34, 100394. <https://doi.org/10.1016/j.edurev.2021.100394>
- Olszewski, B., & Crompton, H. (2020). Educational technology conditions to support the development of digital age skills. *Computers & Education*, 150, 103849. <https://doi.org/10.1016/j.compedu.2020.103849>
- Parkes, J., & Harris, M. B. (2002). The Purposes of a syllabus. *College Teaching*, 50(2), 55–61. <https://doi.org/10.1080/87567550209595875>
- Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11(5), 5-16. <http://doi.org/10.22215/timreview/1438>
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), 33-36.
- Salto-Rivas, R., Novoa-Hernández, P., & Serrano Rodríguez, R. (2022). How reliable and valid are the evaluations of digital competence in higher education: A Systematic mapping study. *SAGE Open*, 12(1), 21582440211068492. <https://doi.org/10.1177/21582440211068492>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504.
- Spante, M., Hashemi, S. S., Lundin, M., & Algers, A. (2018). Digital competence and digital literacy in higher education research: Systematic review of concept use. *Cogent Education*, 5(1), 1519143. <https://doi.org/10.1080/2331186X.2018.1519143>
- Starkey, L. (2020). A Review of research exploring teacher preparation for the digital age. *Cambridge Journal of Education*, 50(1), 37–56. <https://doi.org/10.1080/0305764X.2019.1625867>
- Sun, H., & Ni, W. (2022). Design and application of an AI-based text content moderation system. *Scientific Programming*, 2022. <https://doi.org/10.1155/2022/2576535>
- Thompson, B. (2007). The Syllabus as a Communication Document: Constructing and Presenting the Syllabus. *Communication Education*, 56(1), 54–71. <https://doi.org/10.1080/03634520601011575>
- Tomczyk, Ł., Potyrała, K., Włoch, A., Wnęk-Gozdek, J., & Demeshkant, N. (2020). Evaluation of the functionality of a new e-learning platform vs. previous experiences in e-learning and the self-assessment of own digital literacy. *Sustainability*, 12(23), 10219. <https://doi.org/10.3390/su122310219>
- Wang, X., Wang, Z., Wang, Q., Chen, W., & Pi, Z. (2021). Supporting digitally enhanced learning through measurement in higher education: Development and validation of a university students' digital competence scale. *Journal of Computer Assisted Learning*, 37(4), 1063-1076.
- Weber, H., Hillmert, S., & Rott, K. J. (2018). Can digital information literacy among undergraduates be improved? Evidence from an experimental study. *Teaching in Higher Education*, 23(8), 909–926. <https://doi.org/10.1080/13562517.2018.1449740>
- Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1), 48-59.
- Yang, S. J. H. (2021). Guest Editorial: Precision education – A New challenge for AI in education. *Educational Technology & Society*, 24(1), 105-108.
- Yang, A. C., Chen, I. Y., Flanagan, B., & Ogata, H. (2021a). From human grading to machine grading. *Educational Technology & Society*, 24(1), 164-175.
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021b). Human-centered artificial intelligence in education: seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Yasukawa, M., Yokouchi, H., & Yamazaki, K. (2020). Syllabus mining for analysis of searchable information. *International Journal of Institutional Research and Management*, 4(1), 46-65.
- Yu, S., Su, J., & Luo, D. (2019). Improving BERT-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7, 176600–176612. <https://doi.org/10.1109/ACCESS.2019.2953990>
- Zhao, Y., Pinto Llorente, A. M., & Sánchez Gómez, M. C. (2021). Digital competence in higher education research: A Systematic literature review. *Computers & Education*, 168, 104212. <https://doi.org/10.1016/j.compedu.2021.104212>