frontiers | Research Topics

Artificial intelligence for education

Edited by
Mario Allegra, Manuel Gentile, Giuseppe Città,
Frank Dignum and Iza Marfisi-Schottman

Published in
Frontiers in Education
Frontiers in Artificial Intelligence
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-3981-1
DOI 10.3389/978-2-8325-3981-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Artificial intelligence for education

Topic editors

Mario Allegra — Institute for Educational Technology - National Research Council of Italy, Italy

Manuel Gentile — Institute for Educational Technology - National Research Council of Italy, Italy

Giuseppe Città — Institute for Educational Technology of the National Research Council of Italy (CNR), Italy

Frank Dignum — Umeå University, Sweden

Iza Marfisi-Schottman — EA4023 Laboratoire d'Informatique de l'Université du Mans (LIUM), France

Citation

Allegra, M., Gentile, M., Città, G., Dignum, F., Marfisi-Schottman, I., eds. (2023). *Artificial intelligence for education*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-3981-1

Table of contents

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 07 November 2023
DOI 10.3389/feduc.2023.1276546

OPEN ACCESS

EDITED BY
Eileen Scanlon,
The Open University, United Kingdom

REVIEWED BY
Simon Buckingham Shum,
University of Technology Sydney, Australia

*CORRESPONDENCE
Giuseppe Città
✉ giuseppe.citta@itd.cnr.it

RECEIVED 12 August 2023
ACCEPTED 25 October 2023
PUBLISHED 07 November 2023

CITATION
Gentile M, Città G, Marfisi-Schottman I, Dignum F and Allegra M (2023) Editorial: Artificial intelligence for education. *Front. Educ.* 8:1276546. doi: 10.3389/feduc.2023.1276546

Editorial: Artificial intelligence for education

Manuel Gentile[1], Giuseppe Città[1]*, Iza Marfisi-Schottman[2], Frank Dignum[3] and Mario Allegra[1]

[1]Institute for Educational Technology, National Research Council of Italy, Palermo, Italy, [2]EA4023 Laboratoire d'Informatique de l'Université du Mans (LIUM), Le Mans, France, [3]Department of Computing Science, Umeå University, Umeå, Sweden

KEYWORDS

Artificial Intelligence and Education (AIED), education, generative AI, learning processes, intelligent tutor systems

Editorial on the Research Topic
Artificial intelligence for education

When the Research Topic "*Artificial intelligence for education*" was launched in June 2021, the impact that advances in artificial intelligence would have on the education sector was not entirely predictable.

However, the long and close relationship between research in the two fields of AI and Education was common knowledge. Indeed, since understanding how people learn is closely related to the idea of intelligence, or given that knowledge representation has been one of the most prominent Research Topics in AI, a natural connection between the areas of knowledge concerning Artificial Intelligence and Education emerged even before the term "Artificial Intelligence" was coined (Turing, 1950).

Scholars in the field of artificial intelligence have always looked to the field of education as one of their favorite application areas. From the realization of the logic theorist (Newell and Simon, 1956) to the emergence in the 1990s of cognitive architectures (Laird et al., 1987; Newell, 1990), many of the innovations in the field of AI have found a direct application in the field of education, in the realization of tools such as expert systems to support learning processes and intelligent tutor systems (Anderson et al., 1985; Bidarra et al., 2020).

The new renaissance of AI, marked by recent innovations in the field of deep learning, has in recent years outlined a landscape in which a strong impact could also be expected in education. However, the disruption caused by the market introduction of ChatGPT in November 2022 coincided with the final part of the call for papers for this Research Topic. This timing has therefore cut off, from many of the studies presented in this Research Topic, all the latest research, especially that related to generative AI and large language models (LLM).

Nevertheless, the topic we have been supervising for the past 2 years has allowed us to closely monitor this rapid change, collecting contributions that have proposed and analyzed various topics related to AI in Education. Two recent contributions to the topic by Mallik and Gangopadhyay and Gentile et al. provide an overview of the trend.

Mallik and Gangopadhyay examine how AI, machine learning and deep learning methods are currently used to support the educational process. They conduct this examination by analyzing the involvement of AI-driven methods in the educational process considered as a whole. Based on the analysis of a large set of papers, the authors outline the main trends of future research concerning the use of AI in Education with particular reference to some paradigmatic shifts in the approaches analyzed.

Gentile et al. analyze one of the most exciting topics about AI and Education: the impact of AI on teachers' roles through a systematic literature review. Teachers have always been called upon to change their practices by attempting to integrate new technologies rather than rejecting them. However, even at first glance, the potential changes introduced by AI signal a radical change, what can be called a genuine paradigm shift in teachers' role in Education. According to the authors, the literature analysis reveals that full awareness of the urgency with which the challenges imposed by AI in Education must be addressed has yet to be achieved. Moreover, the study proposes a manifesto to guide the evolution of teachers' roles according to the paradigm shift proposed by Kuhn in the scientific field.

To be managed adequately and avoid causing discomfort in education systems, the assumed changes in the teacher's role should be accompanied by appropriate professional development programmes. In this regard, Sáiz-Manzanares et al. address the topic of designing teacher training programmes that combine the use of technology and instructional design to promote the development of Self-Regulated Learning and automatic feedback systems. Through a study involving 23 secondary school teachers in a training programme delivered with Moodle, the authors investigated the differences in the behavior of experienced and inexperienced teachers, the consistency of the behavior patterns extracted during the study, with the respective type of teacher being modeled, and the teachers' level of satisfaction with the training activity on digital didactics.

The development of assessment tools is one of the preferred areas of application of AI in Education, and, in this respect, AI-based learning analytics will play a key role.

Student-generated texts represent an essential but often unexplored source of information for gaining deeper insights into learners' cognition and ensuring better compliance with students' real needs. To this regard, Berding et al. present a new approach based on applying item response theory concepts to content analysis for the analysis of the textual data generated by the student. They present the results of three studies conducted to make textual information usable in the context of learning analytics. By producing a new content analysis measure, simulating a content analysis process and analyzing the performance of different AI approaches for interpreting textual data, they show that AI can reliably interpret textual information for learning purposes and also provide recommendations for an optimal configuration of AI.

Fleckenstein et al. present a systematic review to explore the effectiveness of AI-based Automated writing evaluation (AWE) tools in realizing systems capable of assessing students' writing skills and providing them with timely feedback with a view to formative assessment. The results confirm a medium-size effect and highlight how it is necessary to continue the exploration by identifying groups of interventions that are more homogeneous among themselves, trying to identify those factors that distinguish these interventions.

Cloude et al. propose an analysis and interpretation framework of real-time multimodal data to support students' Self Regulated Learning (SRL) processes. Specifically, their paper thematises the issues researchers and instructors face when using the data collected through innovative technologies. By recalling a specific procedure through which a researcher/instructor can standardize, process, analyze, recognize and conceptualize multimodal data, they discuss various implications for constructing valid and effective AI algorithms to foster students' SRL.

Cheng et al. address the topic of personalisation of learning using dynamic learning data to track the state of students' knowledge over time. Specifically, the authors propose a context-aware attentive knowledge query network model that can combine flexible neural network models with interpretable model components inspired by psychometric theory to analyze the exercise data.

Chichekian and Benteux propose an exploratory review to describe how the effectiveness of AI-based technologies is measured, the roles attributed to teachers and both theoretical and practical contributions. From the research conducted, it emerges, according to the authors, that the role of teachers is underestimated and that the optimisation of AI systems is still nested exclusively in a strictly IT perspective.

The conscious and informed use of AI and tools that make use of AI is a critical indicator of the maturity of the community that benefits from these instruments. On the contrary, conscious use allows all the potential that can be found in AI to be turned into concrete gains. In this regard, Zammit et al. emphasize the importance of the diffusion and understanding of AI and Machine Learning and the associated ethical implications. To this end, the authors exploit a digital game designed and developed to teach AI and ML core concepts and to promote critical thinking about their functionalities and shortcomings in everyday life.

The paper by Ninaus and Sailer also fits into the groove of critical and aware use of AI in Education. The authors explore humans' role in decision-making in designing and implementing artificial intelligence in Education. Considering the essential role of users in decision-making in educational contexts and emphasizing the need to balance human- and AI-driven decision-making and mutual monitoring, they address both cases in which some AI implementations might make decisions autonomously and cases in which students and teachers, having received information from an AI, are enabled to make reasoned decisions.

Much remains to be done to understand how AI is changing educational practices and how the key stakeholders in the educational community (i.e., students, teachers, faculty, and families) perceive this ongoing change. Nevertheless, the Research Topic provides a broad picture of ongoing changes and a starting point in a research path that will develop over the coming years involving many experts in AI and Education fields.

We believe it is important to renew this Research Topic so that the most recent findings can be shared and systematically analyzed in order to support the progress of this field.

Author contributions

MG: Writing—original draft, Writing—review & editing, Conceptualization, Validation. GC: Writing—original draft, Writing—review & editing, Validation. IM-S: Validation, Writing—review & editing. FD: Validation, Writing—review & editing. MA: Validation, Writing—review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, J. R., Boyle, C. F., and Reiser, B. J. (1985). Intelligent tutoring systems. *Science* 228, 456–462. doi: 10.1126/science.228.4698.456

Bidarra, J., Simonsen, H. K., and Holmes, W. (2020). "Artificial Intelligence in Teaching (AIT): A road map for future developments," in *Empower EADTU, Webinar week: Artificial Intelligence in Online Education.* doi: 10.13140/RG.2.2.25824.51207

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: an architecture for general intelligence. *Artif. Intell.* 33, 1–64. doi: 10.1016/0004-3702(87)90050-6

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A. and Simon, H. (1956). The logic theory machine-a complex information processing system. *IEEE Trans. Inform. Theory* 2, 61–79. doi: 10.1109/TIT.1956.1056797

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 49, 433–460. doi: 10.1093/mind/LIX.236.433

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 March 2022
doi: 10.3389/fpsyg.2022.776806

Teacher Training Effectiveness in Self-Regulation in Virtual Environments

María Consuelo Sáiz-Manzanares[1], Leandro S. Almeida[2], Luis J. Martín-Antón[3], Miguel A. Carbonero[3] and Juan A. Valdivieso-Burón[3]*

[1]Departamento de Ciencias de la Salud, Facultad de Ciencias de la Salud, Research Group DATAHES, Universidad de Burgos, Burgos, Spain, [2]Instituto de Educação, Research Group CIEd, Universidade do Minho, Braga, Portugal, [3]Department of Psychology, Excellence Research Group GR179 Educational Psychology, University of Valladolid, Valladolid, Spain

OPEN ACCESS

Edited by:
Giuseppe Città,
Istituto per le Tecnologie Didattiche ITD - Consiglio Nazionale delle Ricerche, Italy

Reviewed by:
Cesar Collazos,
University of Cauca, Colombia
Christos Troussas,
University of West Attica, Greece

***Correspondence:**
María Consuelo Sáiz-Manzanares
mcsmanzanares@ubu.es

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: 14 September 2021
Accepted: 23 February 2022
Published: 28 March 2022

Citation:
Sáiz-Manzanares MC, Almeida LS, Martín-Antón LJ, Carbonero MA and Valdivieso-Burón JA (2022) Teacher Training Effectiveness in Self-Regulation in Virtual Environments. Front. Psychol. 13:776806. doi: 10.3389/fpsyg.2022.776806

Higher education in the 21st century faces the challenge of changing the way in which knowledge is conveyed and how teachers and students interact in the teaching-learning process. The current pandemic caused by SARS-CoV-2 has hastened the need to face up to this challenge and has furthered the need to approach the issue from the perspective of digitalisation. To achieve this, it is necessary to design training programmes geared towards teaching staff and which address both the use of technology and instructional design aimed at promoting the development of self-regulated learning (SRL) and automatic feedback systems. In this study, work was carried out with 23 teachers (8 inexperienced and 15 experienced teachers) in a training programme conducted through Moodle. The aims were: (1) to test whether there were any significant differences between the behaviour patterns of new teachers compared to experienced teachers, (2) to determine whether clusters of behaviour patterns corresponded to the type of teacher and (3) to ascertain whether the level of teacher satisfaction with the training activity in digital teaching will depend on the type of teacher. A quantitative as well as a qualitative design was applied. Differences were found in the behaviour patterns in the training activities for the development of rubrics and use of learning analytics systems in virtual learning environments. It was also found that the type of teacher did not correspond exactly to the behaviour cluster in the learning platform. In addition, no significant differences were found in the level of satisfaction between the two kinds of teacher. The main contribution this study makes is to provide a detailed description of the training stage as well as the materials required for its repetition. Further analytical studies are required on teacher perception of training programmes in digital teaching in order to provide personalised training proposals that lead to an effective use of teaching in digital environments.

Keywords: self-regulated learning, gamification, learning management systems, virtual environments, teacher training, higher education

INTRODUCTION

Self-Regulation in Higher Education

Recent changes in higher education reinforce students' active role in their learning and skills development. Students' characteristics in terms of academic background, capacities and motivation are assumed as relevant variables in teaching planning; particularly in the case of first year students. Internationally, the literature points to high levels of underachievement and dropout rates for first year students (Bernardo et al., 2017; Páramo Fernández et al., 2017), which can be related to the fact that students commence their higher education studies with little knowledge and few skills in learning strategies or with little information about how to learn new curricular content (Kramarski and Michalsky, 2009).

If they are to ensure an autonomous and active role, students need appropriate levels of autonomy or self-regulation strategies in their learning. Zimmerman (2008) identifies three basic moments in learning self-regulation: planning, performance (monitoring) and self-evaluation. During these phases, an ensemble of thoughts, feelings and actions can be planned, implemented and adjusted by students to improve motivation, learning and achievement (Zimmerman, 2008; Zeynali et al., 2019). It is also important to regulate emotions (Pekrun et al., 2011) in order to achieve optimal performance.

Planning encompasses cognitive processes, prior knowledge, frequent habits and behaviours, as well as motivation and initial expectations. Two processes converge in this first phase: task analysis and demands, and expectations and self-efficacy perceptions (Boekaerts and Niemivirta, 2000). The main impact of good planning translates to an appropriate definition of goals and outlines the strategic plan required to achieve them (Zimmerman, 2013). Performance or execution monitoring is related to what occurs during learning; for example, levels of motivation, attention and self-monitoring (Schunk and Ertmer, 2000; Weinstein and Acee, 2018). These are clearly decisive processes in terms of learning quality and learning outcomes; in other words, with regard to the internal or external feedback that students can receive during task execution (Cervone, 1993; Schunk, 1995; Rheinberg et al., 2000; Kubik et al., 2021). Finally, self-evaluation occurs after task completion and after the achievement obtained has been analysed. Good self-regulation skills enable students to balance initial objectives and learning outcomes, to review the directions taken and the choices made, to consider contextual variables and to take into account all these variables in order to evaluate outcomes or performance and so produce self-evaluation, self-reinforcement and causal attributions (Bandura, 1986; Schunk, 1996; Zimmerman, 2000).

Self-regulation is a complex construct and authors recognise its multidimensionality. Instruments to evaluate self-regulation strategies or skills usually integrate the domain of basic knowledge, cognitive, metacognitive, emotional and motivational student resources (Zimmerman, 2000; Zimmerman and Schunk, 2011). In a contextual approach, self-regulation includes not only traditional cognitive and motivational factors but also regulation of emotions (Calkins and Williford, 2009; Raftery and Bizer, 2009; McClelland et al., 2010; Pekrun et al., 2011; Liew, 2012), the domain of specific knowledge and the level of use of electronic equipment and information. In addition, in terms of cognitive and metacognitive components, authors now pay greater attention to learning strategies and approaches, working memory, inhibitory control or thinking flexibility rather than to classical intelligence or IQ (Carlson, 2003; Rothbart et al., 2011; McClelland et al., 2014; Valadas et al., 2017).

Self-regulation strategies are no doubt related to other student characteristics but are also dependent upon teachers' teaching and evaluation practices. Curricula plans in different degrees can also be an important moderating variable in student self-regulation development. Several programmes are usually introduced in an effort to promote these skills, particularly self-regulation. Institutions and teachers might need to implement diagnostic techniques to identify those skills which are most absent (cognitive, metacognitive, motivational and emotional), dealing with specific student subgroups.

Advanced Learning Technologies and Self-Regulated Learning

The use of technology and educational data mining techniques (EDM) form part of the Advanced Learning Technologies (ALT) methodology. ALT is triggering a revolution in the field of cognitive psychology and learning, since it facilitates both the development and evaluation of the teaching-learning process. Much of today's learning is carried out in virtual spaces. These environments aid self-regulated learning (SRL; Azevedo et al., 2011, 2015) through a range of different virtual reality resources and hypermedia, such as avatars and serious games (Kretschmer and Terharen, 2019; Sáiz-Manzanares et al., 2020). Van De Weijer et al. (2020) found that the use of gamification enhances students' cognitive skills and boosts motivation (Nappo et al., 2020) in high duration interventions (24 weeks). The use of executive functions (control and self-regulation) is particularly important vis-à-vis acquiring new concepts or learning that involves a high degree of difficulty. These skills are directly related to establishing goals and to planning, and acquiring these skills is linked to achieving successful educational responses (Huizinga et al., 2018). Implementing metacognitive strategies can be enhanced through the use of serious games. Nevertheless, such interaction entails the need to have experts in learning psychology, in the development of virtual environments as well as experts in artificial intelligence, since analysing the results of platform learners will provide insights into and shed light on what the most appropriate type of game is for each user. As a result, gamification emerges as a help in the more efficient use of executive functions (attention, inhibition of distracting elements, planning and self-evaluation) as well as increased motivation. Specifically, the use of gamified learning strategies within virtual learning environments (Learning Management Systems -LMS-) enhances the quality of learning and engenders greater student motivation compared to conventional forms of learning (Pinnell, 2015). Moreover, the value of the effect within the differences found ranges between $d = 0.45$—$d = 0.72$, implying a medium-high effect (Taub et al., 2018). This appears to be because these activities help information

to be processed in the working memory and in the long-term memory and prevent task execution from being abandoned (Lumsden et al., 2017).

Moreover, the joint use of LMS and ALT enables interactions to be recorded (Azevedo and Gašević, 2019; Hosain et al., 2019; Noroozi et al., 2019). The use thereof accounts for over 72% of variance in student learning outcomes (Sáiz-Manzanares et al., 2019a). One possible reason is that the use of ALT boosts SRL learning and the use of metacognitive strategies (planning, evaluation and design of task solving; Hull et al., 2015) as well as student motivation (Zimmerman, 2005), all of which enhances personalised learning (Enembreck and Barthès, 2005; Sáiz-Manzanares et al., 2019b; Martín-Antón et al., 2020), learner autonomy (Remesal et al., 2017; Zorrilla-Pantaleón et al., 2021) as well as self-awareness and self-reflection (Taub et al., 2017; Nurmi et al., 2020).

Nevertheless, research is required into the design of such environments, since the mere use of virtual platforms by no means ensures effective learning (Yamada and Hirakawa, 2016; Park and Jo, 2017; Sáiz-Manzanares et al., 2017). Carefully designed methodological aspects (objectives, conceptual and procedural content, assessment criteria) as well as technological aspects (Sáiz-Manzanares et al., 2019a) must be applied if these environments are to foster the development of metacognitive strategies and self-regulation. Moreover, virtual learning platforms must embrace student follow-up systems so that teachers can track the behaviour of each of their students throughout the learning process (Jommanop and Mekruksavanich, 2019; Troussas et al., 2021; Krouska et al., 2021b; Sáiz-Manzanares et al., 2021b).

Teacher Training in Higher Education in Effective Teaching in Virtual Environments

As has become clear through the previous points, the teaching-learning process in LMS, particularly in higher education contexts, involves addressing digital transformation. This has been hastened by the current situation triggered by the SARS-CoV2 pandemic (García-Peñalvo, 2021; Sáiz-Manzanares et al., 2021a). Said crisis is having an impact on the teaching-learning process, particularly in higher education, since it is leading to a situation of uncertainty which is reflected in emotional behaviour related to anxiety during the process, both amongst teachers and students alike (de la Fuente et al., 2021a). This prompts the need to develop teaching models based on preventing the situations of uncertainty that trigger anxiety (de la Fuente et al., 2021b). In order to meet the challenge of a true digital transformation in higher education, technological resources together with innovation in teaching processes must be introduced (García-Peñalvo and Corell, 2020). All of this leads to teacher training, which will need to focus on content handling of LMS and ALT resources (e.g., avatars, gamification and automatic feedback procedures). This challenge in terms of training is one of the goals of government authorities included in objective 5, quality of teaching, of the 2030 Agenda (Redecker and Punie, 2017; Jarillo et al., 2019). In this line, the European Commission has established a Framework for the Digital Competence of Educators (DigCompEdu; Redecker and Punie, 2017). DigCompEdu defines six levels of teaching staff competence: (A1) Newcomers (teachers who have had very little contact with digital tools); (A2) Explorers (teachers who have begun to use digital tools, but who lack a global or consistent approach, such that they need to expand their skills); (B1) Integrators (use and experiment with digital tools for a variety of purposes, seeking to determine which digital strategies function best in each context); (B2) Experts (use a range of digital tools with confidence, creativity and a critical eye to improve their professional activities. They are constantly expanding their repertoire of practical work); (C1) Leaders (use a wide range of flexible, comprehensive and effective digital strategies. They are a source of inspiration for other teachers); (C2) Pioneers (question the suitability of the contemporary digital and pedagogical practices which they themselves are experts in. They lead the way in innovation and are a model for younger teachers). The ultimate goal is to train professionals with skills in educational digitalisation in order to increase motivation and help students achieve efficient learning (Carbonero et al., 2017).

Moreover, training in digital skills amongst teachers, particularly within the framework of higher education, is a challenge that requires implementing formal training programmes (García-Peñalvo, 2021). The content of these training proposals in e-Learning or b-Learning spaces during the COVID-19 pandemic in higher education must take into account (Collazos et al., 2021; de la Fuente et al., 2021b):

- Frequent interaction through technological resources at specific times (scheduled synchronised sessions).
- Expectations of normality in work during the teaching-learning process.
- Fostering collaborative work and assessment systems with feedback on the process.
- Facilitating SRL through technological resources in LMS.
- Incorporating personalised consultation (through videoconferences, forums or chats).
- Safeguarding students' emotional state, avoiding loneliness in the net.

In addition, gaining an insight into teachers' perception of the educational processes related to the use of technological resources in teaching as well as distinguishing between inexperienced and experienced teachers is key to improving teaching processes in today's society, particularly given the current worldwide pandemic (Krouska et al., 2020). Moreover, it is important to develop teaching models that take into account the emotional and social aspects of cognitive and metacognitive development within the framework of e-learning or b-learning, which is undoubtedly here to stay (Dumulescu et al., 2021). Furthermore, designing these learning environments is key to the success of the teaching-learning process (Collazos et al., in press).

Taking into account the conclusions to emerge from the previously mentioned studies, the research questions (RQ) for the study were:

1. "Will the behaviour patterns of university teachers during a training activity in digitalisation in Moodle depend on whether they are inexperienced or experienced teachers?"
2. "Will behaviour clusters in LMS correspond to the differentiation between the type of teacher (inexperienced or experienced)?"
3. "Will the level of satisfaction with the training activity in digital teaching depend on the type of teacher (inexperienced or experienced)?"

This study applied mixed methods, merging quantitative and qualitative analyses (Anguera, 1986; Castañer-Balcells et al., 2013). Specifically, a quantitative and qualitative study was used to test RQ1 and RQ2, and a quantitative study was carried out to test RQ2.

MATERIALS AND METHODS

Participants

We worked with a total sample of 23 teachers, 15 experienced teachers (with over 15 years teaching in higher education), nine females and six males, and 8 inexperienced teachers (with 1–2 years teaching experience in higher education), seven female and one male, from four universities (University of Burgos, University of Oviedo, University of Minho and University of Valladolid). Experienced teachers were aged between 45 and 60, and inexperienced teachers were aged between 25 and 30. Prior to commencing the study, all the participants were informed of the aim of the research and their written consent was requested. A convenience sample was used to select participants. Participants were selected by each partner involved in the SmartArt project, following the guidelines set out in the project report a learning activity is organised for two students and two teachers for each partner (eight students and eight teachers in all) chosen at random from amongst the participating organisations. However, the number of participants was increased depending on the requests put forward by each partner. Throughout the study, 2 experimental deaths were detected in the group of experienced teachers.

Instruments

Initial Survey on Prior Knowledge of ALT

An *ad hoc* survey was drawn up to ascertain participants' level of prior knowledge of the training activity related to their know-how and application of teaching resources in virtual learning environments (Sáiz-Manzanares, 2021). The survey consisted of nine closed response questions, measured on a 1–5 Likert-type scale, with 1 being the lowest level of prior knowledge and 5 the highest. Survey reliability was determined by applying the composite reliability index, Omega index, with the value for the general scale being $\Omega = 0.90$. Two open response questions were also included (1. What are your expectations towards the training activity? What would you like to learn in the training activity?). This survey is available in **Supplementary Table S1**.

Application Web UBUMonitor

UBUMonitor is an open-code and free computer application (Ji et al., 2018). The application runs in the client and is implemented through Java, and it has a graphic interphase developed in JavaFX. The application is connected with the chosen Moodle server through web services and the API REST provided by the server. When no web services are available to retrieve specific data, web scraping techniques are also used. All the communication between the Moodle server and the client UBUMonitor is encrypted *via* HTTPS for security reasons. As a result of these queries, data are obtained in JSON and CSV format, processed and transformed into Java objects in the client. Java and webpages are applied with different graphic libraries of JavaScript within the desktop application in order to visualise the data gathered. The application includes six modules: visualisation (which offers frequency representation in different graphics: Heat Map, Boxplot, Violin, Scatter, etc.), comparisons, forums, dropout rate risk (locating students who have failed to log on for 7–15 days at certain moments of the course), Calendar of events and Clustering (finding clusters by applying different algorithms such as *k*-means ++, Fuzzy *k*-means, etc.). Specifically, in this study we used the visualisation module, which allows for an analysis of access frequency in components, events, sections or courses seen in Moodle, with options to analyse the registers in different graphics. In this work, we opted for the Heat Map visualisation technique, since it provides the results with numerical and colour intensity visualisation throughout the course during the training activity. The use of visualisation techniques such as Heat Map is felt to be very useful to assess user behaviour in LMS (Dobashi et al., 2019). The UBUMonitor application may be downloaded free at https://github.com/yjx0003/UBUMonitor.

Training Programme for University Teachers

This programme was implemented in the LMS based on Moodle UBUVirtual. It was also based on the use of ALT, grounded in the use of gamification in self-assessment systems to promote SRL. The training programme lasted 4 weeks and consisted of a synchronised online phase made up of five 3-h training sessions. These sessions were carried out in UBUVirtual through the joint communication and collaboration Teams platform. A description of the phases of the synchronised sessions can be seen in **Figure 1**. The documents related to the teaching staff training sessions may be consulted at: https://bit.ly/3vsS94l.

Each of the sessions had a consistent pedagogical structure comprising presentations on the topics dealt with during each session: a collaborative work chat to deal with doubts, complementary documentation, gamification activities to understand the concepts of the topic and a satisfaction survey for the training activity. **Figure 2** sets out the structure. Training in all the sessions was offered in Spanish and in English. The specific content of the synchronised training sessions may be consulted in **Supplementary Table S2**. This training structure follows the approach of acquiring executive control strategies, since these initially seek to focus participant attention on the content to be dealt with in each unit. They then direct planning

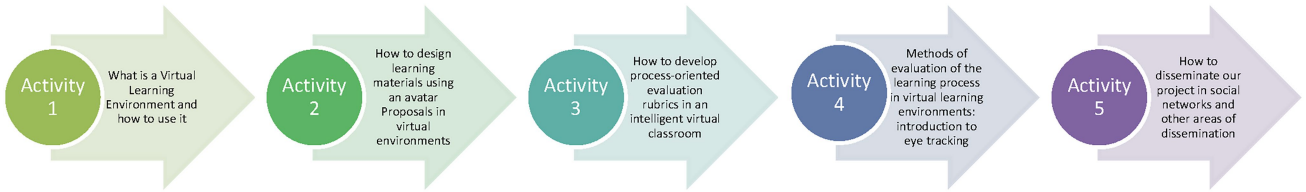


FIGURE 1 | Structure of the training in synchronised phases.

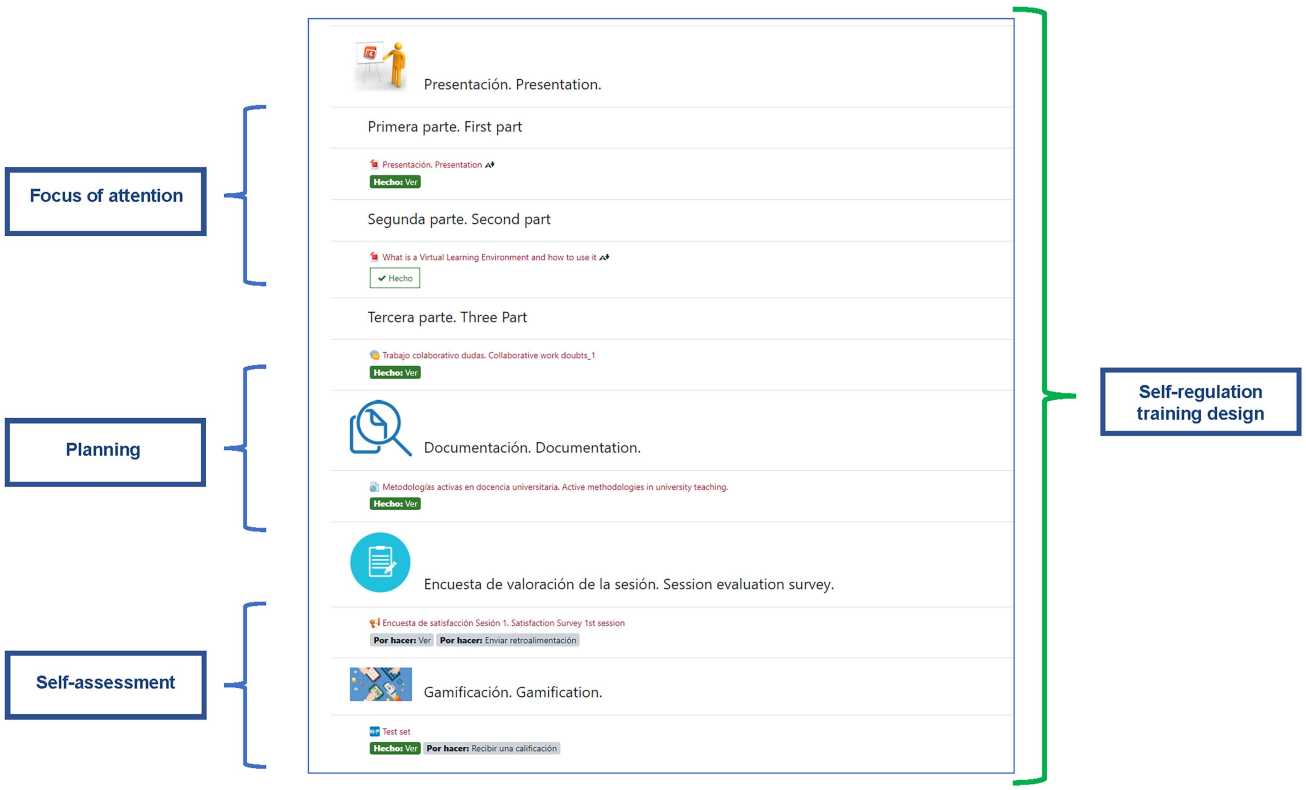


FIGURE 2 | Didactic structure of each of the training sessions.

strategies in order to establish the learning goals related to the content. Finally, they focus on the acquisition of self-evaluation strategies, in this case through gamified learning techniques with automatic feedback and with satisfaction surveys that encourage reflection on the learning process.

The gamification activities designed for each training session can be seen in **Supplementary Table S3**. All of them were designed using the HTML5 package (H5P). H5P is a totally free and open technology, with an MIT licence. Information may be found at https://h5p.org/. H5P is a resource that may be implemented in LMS similar to Moodle, WordPress or Drupal and which enables educators to create different types of content. The following resources were specifically used in this study: Drag the Words (allowing challenges to be created based on text in which users have to drag words into gaps in the sentences), Find the words (users have to find a series of keywords in the grid) and Multiple Choice (multiple choice questions). It also includes instant feedback on the correct options and the reasons for these and True/False Question (refers to true-false questions). All of these serious games involved feedback on the answers as well as information on progress.

The training programme also involved a synchronised training phase that took place over a three-week period. During this phase, teachers had to develop a teaching design proposal for each university group to be applied in a virtual learning environment. This design had to include one of the tools seen during the synchronised training phase. Interaction was by email or through a forum set up for the purpose on the UBUVirtual platform. This training design was similar in structure to the one which teachers would be expected to include during their teaching in higher education.

Satisfaction Survey With the Synchronised Sessions

An *ad hoc* survey was designed to gauge participant teacher satisfaction with the synchronised training activity. The survey was made up of four closed response questions measured on a 1 to 5 Likert-type scale, where 1 reflects the lowest level of satisfaction and 5 the highest, in which satisfaction is assessed with the concepts, materials, complementary information and work time devoted to the activity, together with three open questions [(1) indicate which aspects need to be extended in this part of the course, (2) indicate the aspects to be removed from this part of the course, and (3) suggestions for improvement]. Survey reliability was attained by applying the composite reliability index, Omega, and which gave $\boldsymbol{\Omega} = 0.62$. This instrument is available in **Supplementary Table S4**.

Satisfaction Survey With the Training Activity

This survey was designed *ad hoc* and was based on the assessment criteria of the European Commission for the Evaluation of Learning Activities in European projects. The survey is made up of 14 closed response questions, measured on a 1–5 Likert-type scale, where 1 reflects the lowest level of satisfaction and 5 the highest. Survey reliability was attained by applying the composite reliability index, which gave Omega, $\boldsymbol{\Omega} = 0.96$.

The survey also included two open response questions [(1) which of the gamification materials have you found most useful for understanding the concepts? and (2) what elements would you introduce or increase in gamification materials?]. This instrument is available in **Supplementary Table S5**.

Procedure

This research was carried out as part of the "Self-Regulated Learning in SmartArt (SmartArt)" project funded by the European Commission. The aims of the project focus on designing SRL-based virtual intelligent classrooms and the use of avatars to facilitate personalised and independent student learning. For further information, see https://srlsmartart.eu/en.

The project was backed by a favourable report issued by the University of Burgos Bioethical Committee, No. IR 27/2019, the coordinating university. The project was to contain a training phase aimed at university teachers from partner universities and which dealt with teaching strategies in virtual learning platforms based on self-regulated learning through the use of technological resources.

Prior to commencing the study, participating teachers' level of prior knowledge in digital teaching was evaluated. To this end, an *ad hoc* survey was designed—see instruments section. The online training stage, consisting of a synchronised phase (lasting a week), was then carried out. After each synchronised training session, a satisfaction survey was conducted with the synchronised sessions (see "Instruments" section). There was also a non-synchronised phase (lasting 3 weeks). Finally, once the training activity had concluded, participants were given an *ad hoc* satisfaction survey on the activity (see instruments section). A diagram of the procedure used in this study can be seen in **Figure 3**.

Data Analysis

Prior Analysis

Before testing the RQ, a normality study was carried out on the sample, for which asymmetry and kurtosis analyses were applied. The SPSS v.24 statistical package (IBM, 2016) was used for this purpose.

Hypotheses Testing

In order to test the RQ, quantitative and qualitative studies were performed. With regard to the latter, a descriptive design was applied (Campbell and Stanley, 2005), and a comparative longitudinal design was used for the latter (Flick, 2014).

As regards the quantitative study, since some of the asymmetry indicators did not ensure normal distribution and the n of subjects in the sample was below 30, a non-parametric statistic was applied. Specifically, to test RQ1 and RQ3 the Mann–Whitney U test for independent samples was used (Mann and Whitney, 1947), for which the SPSS v.24 statistical package was used (IBM 2016; see Equation (1)).

$$Ui = n1 + n2 + ni\left(ni + 1\right)/2 - Ri$$

where n_1 will be equal to the n of group 1, and n_2 will be equal to the n of group 2, and R_i is construed based on the sum of the ranges of one of the samples chosen at random. The value of the effect size was determined by applying the formula of eta squared [η^2; see Equation (2)]. As regards the interpretation of the values, and following Cohen (1988), a very small effect size was considered to be one between 0 and 20, small between 20 and 49, medium between 50 and 69, with over 70 being considered as high.

$$\eta^2 = \frac{\mathbb{Z}}{N} - 1$$

With regard to testing RQ2, cluster analysis was used, applying the *k*-Means ++ algorithm. This algorithm is applied to select the initial values of the centroids for the *k*-means clustering algorithm. This was proposed in 2007 by Arthur and Vassilvitskii (2007) as an approximation algorithm to address the NP-hard *k*–means problem: in other words, as a way of avoiding the occasionally poor clustering found by the standard *k*-means algorithm [see Equation (3)].

$$D^2\left(\mu_0\right) \leq 2D^2\left(\mu_i\right) + 2\left\|\mu_i - \mu_0\right\|^2$$

being μ_0 the initial point selected and D the distance between point μ_i and the centre closest to the cluster. Having chosen the centroids, the process is like the classical *k*-means. To find this, the UBUMonitor tool was used (Ji et al., 2018).

Also used was Pearson's contingency coefficient C (which expresses the intensity of the relation between two or more qualitative variables, and which is based on comparing the sequences of two characteristics with the expected frequencies). This is calculated by calculating $\chi 2$, adding the categorisations of the

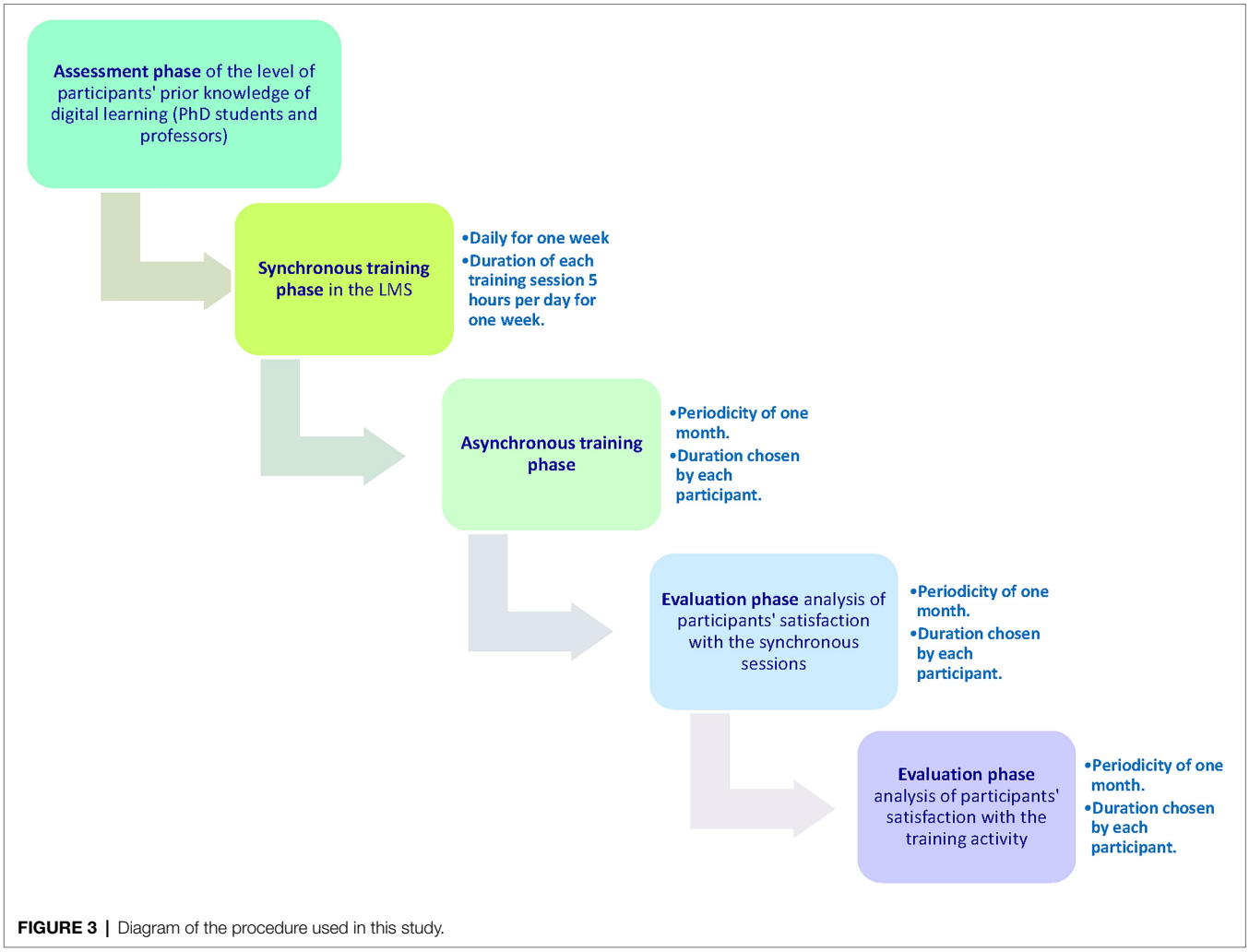


FIGURE 3 | Diagram of the procedure used in this study.

two judges in the analysis of subjects' responses in all the analysis units and then removing empty categories (López-Roldán and Fachelli, 2015; see Equation (4)). The statistical package SPSSv.24 (IBM, 2016) was used to determine this.

$$C = \sqrt{\left(\chi^2 / N + \chi 2\right)}$$

As regards the qualitative study, Heat Map visualisation techniques derived using the UBUMonitor tool (Ji et al., 2018) were used in RQ1, and in RQ3 frequency analysis was used on the categorisation criteria for the open answers to the initial and final evaluation surveys carried out using ATLAS.TI 9 software (Atlas.ti, 2020).

RESULTS

Prior Analysis

Prior to commencing the study, a check was carried out on the distribution of the sample vis-à-vis their previous knowledge in digital teaching. Asymmetry values were adjusted in all the items except in items 4 and 9, in which a slightly higher value was seen (values over |2.00| are considered extreme). As regards the kurtosis values, no extreme values were found (values between |8.00| and |20.00| are considered acceptable; Bandalos and Finney, 2001), see **Table 1**. As a result, a normal distribution was not considered, and a non-parametric statistic was applied to test the research questions.

In the qualitative study of the open response questions, the questions were first categorised and then analysed with the Atlas.ti 9 qualitative analysis program, applying percentage analysis to the categorised responses. Results indicate there were two kinds of interests amongst participating teachers; one part preferred to learn about basic resources for implementing teaching in virtual spaces (33.33%), and another group requested advanced techniques (29.0%). A general interest was also noted in specifically learning about SRL techniques through avatars and gamification techniques (28.57%).

Testing the Research Questions

In order to test RQ1 "Will the behaviour patterns of university teachers during a training activity in digitalisation in Moodle depend on whether they are inexperienced or experienced teachers?"

TABLE 1 | Descriptive statistics and asymmetry and kurtosis values in the initial survey on prior knowledge.

Question	*M(SD)*	*Asymmetry*	*ESA*	*Kurtosis*	*ESC*
1. I believe that the teaching-learning process should be interactive between teacher and student.	5.00(0.00)	-	0.56	–	1.91
2. I have a knowledge of how to design virtual learning platforms.	2.81(1.18)	0.16	0.56	–1.17	1.91
3. I have a knowledge of how to design process-oriented feedback.	3.00(1.17)	0.00	0.56	–0.47	1.91
4. The feedback provided by the teacher on the student's practice should be clear, positive and task-dependent.	4.88(0.33)	–2.51	0.56	4.90	1.91
5. I have a knowledge of eye tracking methodology.	2.44(1.12)	–0.13	0.56	–1.46	1.91
6. I have a knowledge of how to design learning-oriented gamification activities.	3.25(1.30)	–0.33	0.56	–0.99	1.91
7. I have a knowledge of project dissemination in social networks.	2.81(1.07)	0.08	0.56	–0.27	1.91
8. I have previously used gamification experiences as a learning resource.	3.06(1.52)	0.12	0.56	–1.56	1.91
9. I have previously used the Alexa skill to monitor learning activities.	1.44(1.00)	2.28	0.56	3.95	1.91

ESA = Standard Error Skewness.
ESC = Standard Error Kurtosis.

Quantitative Study

An analysis was first carried out to ascertain whether there were significant differences in interaction in the training platform between inexperienced or experienced teachers. In order to test this, we applied the non-parametric Mann–Whitney *U* test of differences between independent samples (see **Table 2**). Two experienced teachers who signed up for the activity later did not take part for personal reasons.

Significant differences were found in platform interaction between inexperienced teachers and experienced teachers in session 3 (designing rubrics in VLE) and a medium effect value ($\eta^2 = 0.50$), session 4 (use of Learning Analytics Systems in VLE) and small effect value ($\eta^2 = 0.46$), in favour of the group of inexperienced teachers and a small effect size ($\eta^2 = 0.46$).

Qualitative Study

In order to analyse RQ1, the heat maps in the various Moodle components were pinpointed, distinguishing between the maps of inexperienced vs. experienced teachers during the synchronised and non-synchronised interaction phases. With regard to behaviour analysis, greater interaction was evident in the UBUVirtual platform during the synchronised phase when compared to the non-synchronised phase for both types of teacher, although interaction frequency was more intense amongst inexperienced teachers (see **Figures 4, 5**).

As regards the analysis of behaviour during the non-synchronised training phase, a decrease was seen in interaction frequency in both types of teacher (inexperienced and experienced), although interaction frequency was greater amongst inexperienced teachers (see **Figures 6, 7**).

In order to test RQ2 "will behaviour clusters in LMS correspond to the differentiation between the type of teacher (inexperienced or experienced)?," we first used an eight-cluster analysis with regard to the number of registers in the platform of the completed activity, for which the *k*-means ++ algorithm was applied (see **Figure 8**).

We then designed a cross-reference table between the allocation cluster and inclusion in the group of inexperienced or experienced teachers (see **Table 3**). We also found the coefficient of contingency, which obtained a value of $C = 0.41$ to be non-significant $p = 0.28$. This indicates there is no strong correspondence between the cluster allocated and the type of group to which the teacher belongs. Non-significance might be due to the small number of elements in the sample.

In order to test RQ3 "Will the level of satisfaction with the training activity in digital teaching depend on the type of teacher (inexperienced or experienced)?" we applied the non-parametric Mann–Whitney U test of differences between independent samples.

An analysis of satisfaction with the overall training activity was performed.

No significant differences were found between the group of inexperienced teachers vs. experienced teachers in the level of satisfaction in any of the items contained in the satisfaction survey for the training activity. The effect value was seen to be low in all the items. Moreover, mean satisfaction scores were high in all the items, with the means interval ranging from 4.33 to 5 out of 5 (see **Table 4**). The significance of the coefficients may be explained by the sample size, which in this study was small.

As regards the analysis of the responses to the open questions, the latter were categorised and analysed with the Atlas.ti 9 qualitative analysis program, applying percentage analysis of the categorised responses. It was found that the training activities that aroused the greatest interest amongst teachers were those related to working with avatars to provide SRL (44.44%) and designing gamification activities to provide student self-evaluation (22.22%). In addition, 73% of teachers considered that the training activity fitted in well with the time and content, with 12.5% indicating that they would

TABLE 2 | Descriptive statistics and Mann–Whitney *U* test comparing teachers (inexperienced or experienced) on the UBUVirtual training platform during the synchronised training phase.

Synchronised training sessions	Session content	Group 1 Inexperienced *n* = 8	Group 2 Experienced *n* = 13	Mann–Whitney *U*	*p*	*Z*	η^2
		M(*SD*)	*M*(*SD*)				
Session 1	Definition and use of Virtual Learning Environments (VLE)	51.75(47.51)	24.76(40.09)	35.00	0.21	−1.25	0.08
Session 2	Design of materials and use of avatars	37.00(43.38)	21.54(43.19)	38.00	0.31	−1.02	0.05
Session 3	Design of rubrics in the VLE	36.00(38.29)	4.00(9.55)	9.00	0.001*	−3.24	0.52
Session 4	Use of Learning Analytics Systems in the VLE	43.25(25.79)	8.54(16.58)	10.50	0.002*	−3.10	0.48
Session 5	Dissemination in social networks	36.00(45.46)	15.85(17.70)	32.00	0.144	−2,32	0.27

*$p < 0.05$.
Z = Z de Kolmogorov–Smirnov; effect size η^2.
M = Mean. SD = Standard Deviation.

FIGURE 4 | Heat Map of inexperienced teacher behaviour in the Moodle platform during the synchronised phase.

reduce slightly the time devoted to the activities during the synchronised phase.

An analysis of satisfaction with each of the synchronised training sessions was also carried out. To do this, the Mann Whitney *U* test was applied to the responses of the satisfaction survey conducted for each synchronised session. The mean satisfaction scores in the evaluation of all the synchronised training sessions were high, since they ranged from 4.06 to 4.82 out of 5 in the group of inexperienced teachers and from 3.79 to 4.80 out of 5 in the group of experienced teachers. Nevertheless, significant differences did emerge in the satisfaction with training session 2 (design of materials and use of avatars), session 4 (use of Learning Analytics Systems in VLE) and session 5 (dissemination in social networks) with regard to the clarity of the concepts explained, in favour of the group of inexperienced teachers (see **Table 5**). In all cases, the effect value was low.

With regard to the analysis of the open response answers, 90% of the teachers would not omit anything, although 10% did indicate that there was a lot of information. As regards suggestions for improvement, 90% felt that there should be more practical training whilst 10% would not add anything.

FIGURE 5 | Heat Map of inexperienced teacher behaviour in the Moodle platform during the non-synchronised phase.

FIGURE 6 | Heat Map of experienced teacher behaviour in the Moodle platform during the synchronised phase.

DISCUSSIONS

With regard to RQ1, it was found that teachers' interaction behaviour pattern in LMS differed depending on whether they were either inexperienced or experienced teachers. In general, inexperienced teachers tended to interact more, both during the synchronised and the non-synchronised phase, although differences in interaction did emerge between the two groups.

FIGURE 7 | Heat Map of experienced teacher behaviour in the Moodle platform during the non-synchronised phase.

These aspects might be related to the teaching style and the internal expectations of teaching staff towards the training activity. Although all the teachers initially started out with the same interest, certain unseen motivations might be exerting an influence. These aspects are related with the results found in the qualitative analysis of the open questions posed in the initial survey, since differences were found in the level of interest displayed towards the training activity. This indicates the need for further inquiry to analyse teaching styles in e-Learning and b-Learning spaces and which explores in depth teachers' internal motivation towards teaching in these spaces. Such an analysis would also examine which factors might account for the differences in interaction found in the synchronised and non-synchronised phases of the training activity amongst the various participants. Following García-Peñalvo (2021), the process of digital transformation within the framework of higher education poses a complex challenge which, if it is to be addressed effectively, requires government training proposals and a micro-analytical analysis of how this training is perceived and applied in real situations.

As regards RQ2, no exact correspondence was found between the type of teacher (inexperienced vs. experienced) and the behaviour patterns displayed in LMS during the training phase. This has also been reported in other studies with university students (Sáiz-Manzanares et al., 2021b), indicating that although they may initially be seen as homogeneous groups, differences do exist that are probably linked with motivation towards training and with the style of learning. This aspect has not been dealt with previously but is now a key reference, since digital transformation demands that teaching staff be trained or that their training be brought up to date.

With regard to RQ3, the motivation of the teachers taking part was found to be very high, regardless of whether they were inexperienced or experienced teachers. Differences did, however, emerge in terms of perception amongst the group of experienced teachers in terms of the following aspects: designing avatars to encourage SRL, use of learning analytics systems in LMS, and use of social networks to disseminate content. This might be explained by the generational difference between inexperienced and experienced teachers in that the former might have a greater degree of digital competence in these aspects.

Although achieving consistency in satisfaction is a complex task, structuring training activities in levels in terms of degree of difficulty and skill acquisition might offer one solution to this issue.

Limitations and Future Lines of Research

The results to emerge should, however, be taken with a certain degree of caution, given the characteristics of the sample (non-random selection, small number of participants and the features thereof—they belong to research groups who are analysing the effectiveness of SRL in the teaching-learning process). Worth highlighting is the need to promote research that applies mixed methods (quantitative and qualitative), since qualitative analysis of the responses to the open questions provides a great deal of information about how the training

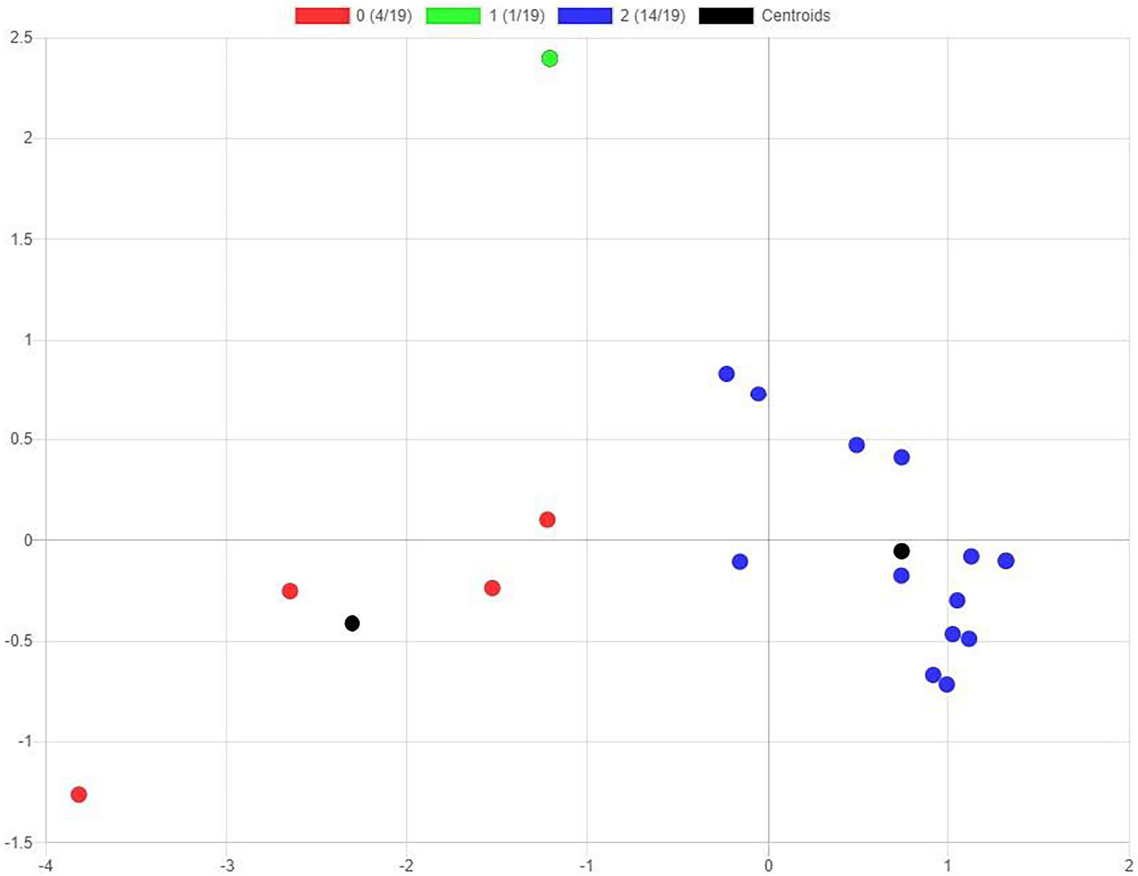


FIGURE 8 | Cluster analysis with the *k*-means ++ algorithm.

TABLE 3 | Cross-reference table between the values of the cluster allocation and the type of teacher; inexperienced vs. experienced.

Type of teacher	Cluster			Total
	0	1	2	
Inexperienced	5	2	1	8
Experienced	10	3	0	13
Total	15	5	1	21

process is perceived and the needs to be pinpointed. This entails carrying out studies that provide for a microanalytical analysis, which in turn means that ratios must not be too big. Future studies will focus on examining what perception teaching staff who evidence different skill levels and who come from different knowledge areas have of training activities in the digitalisation of teaching. It is important to analyse these training activities, given that the current pandemic triggered by COVID-19 the world over means that teaching staff training, which was formerly carried out face-to-face, must now be done online. Furthermore, the actual training content must respond to what is needed in training teaching skills in digital environments. Further research is thus required into how effective these prove to be. What was previously an optional form of training has now become almost the only form such that, although the study does evidence certain limitations, which are mainly related to the generalisation of the results due to the nature of the sample, it does nevertheless afford the advantage of being a study based on the individual follow-up of participants through various monitoring tools. It also offers a detailed list of the materials and tools applied, which can be consulted in the supplementary material and in the open-access links provided, all of which helps with the replicability of the work.

By way of a summary, **Table 6** provides a synopsis of the results found in the studies that served as justification for this work, together with the results to emerge from the work itself.

CONCLUSION

Higher education faces a major challenge in terms of teaching in the 21st century. Said challenge, which had already been set out by government authorities in objective 5 of the 2030 Agenda (Redecker and Punie, 2017), has been hastened as a result of the current health crisis brought on by the pandemic (García-Peñalvo, 2021).

TABLE 4 | Descriptive statistics and Mann–Whitney *U* test for the results in the satisfaction survey for the training activity in participating teachers (inexperienced vs. experienced).

Final evaluation survey on the activity	Group 1 Inexperienced $n = 8$	Group 2 Experienced $n = 13$	*U* Mann–Whitney	*p*	*Z*	η^2
	M(*SD*)	*M*(*SD*)				
1. Communication with the meeting coordinator.	5(0.00)	5(0.00)	24	0.65	0	0.00
2. Learning activity agenda.	4.83(0.41)	4.63(0.74)	21.50	0.65	−0.45	0.01
3. Presentation on the ongoing progress by the project coordinator.	4.67(0.52)	4.63(0.74)	23.00	0.87	−0.16	0.00
4. Time management.	4.50(0.55)	4.63(0.74)	19.50	0.49	−0.69	0.02
5. Atmosphere and communication among attendees.	4.50(1.23)	4.38(1.19)	22.50	0.79	−0.27	0.00
6. Would you be interested in using the tools proposed in your job?	4.67(0.82)	4.38(1.10)	19.50	0.47	−0.73	0.03
7. Do you consider that the tools presented are easy to use when teaching?	4.33(1.21)	3.63(1.30)	14.50	0.20	−1.29	0.08
8. Do you consider that specific training is necessary to use the tools presented?	4.50(0.84)	4.38(0.74)	18.00	0.41	−0.83	0.03
9. Would you like to spread the proposed tools among your colleagues?	4.50(0.84)	4.38(0.74)	21.00	0.66	−0.44	0.01
10. Quality of the virtual environment in which the training action was carried out.	4.67(0.52)	4.63(0.74)	23.00	0.87	−0.16	0.00
11. The gamification activities have made it easier for me to understand the concepts.	4.00(0.89)	4.25(1.04)	19.00	0.49	−0,69	0.02
12. Their satisfaction with the duration of the training activity is.	4.50(0.84)	4.50(0.76)	23.50	0.94	−0.78	0.03

Z = Z de Kolmogorov–Smirnov; effect size η^2.
M = Mean. SD = Standard Deviation.

Implications for Teacher Training in Higher Education

In higher education, face-to-face teaching as the only means of teaching is dying out. Current higher education teaching is delivered through e-Learning or b-Learning (García-Peñalvo, 2021). This implies that instructional design must undergo changes compared to the traditional design. These changes are related to the use of the technological and pedagogical resources afforded by student SRL and self-evaluation in order to provide personalised learning (Jommanop and Mekruksavanich, 2019; Kubik et al., 2021; Sáiz-Manzanares et al., 2021b). This entails the use of tools that offer the student intelligent tutoring in LMS (Azevedo et al., 2011, 2015; Taub et al., 2017, 2018; Troussas et al., 2021; Krouska et al., 2021a). In order to achieve this, two key aspects are required; firstly, the functional pedagogical design of LMS that will allow for the inclusion of technology-based resources such as avatars and gamification activities. These environments must also include easy-to-use follow-up tools for tracking student learning behaviour throughout the teaching-learning process (Jommanop and Mekruksavanich, 2019; Sáiz-Manzanares et al., 2021b; e.g. UBUMonitor).

If the challenges facing teaching within the framework of higher education are to be met successfully, it is necessary to design and implement training programmes (de la Fuente et al., 2021a, 2021b) that skill teachers in the use of LMS and the technological tools included in these virtual environments so as to automatically provide SRL and self-evaluation (Dumulescu et al., 2021; Kubik et al., 2021). These training proposals must offer varying levels of difficulty vis-à-vis the acquisition of skills, adapting to each teacher's training requirements. This prioritisation of skills is related to teachers' previous knowledge in digitalisation and with the style of teaching they have employed throughout their teaching career. It must be borne in mind that experienced teachers have a background of teaching based on face-to-face interaction. This means that although they may have used innovative teaching techniques, they will have done so in face-to-face contexts. The interaction between teacher and student and between students themselves in these spaces differs enormously from what is found in e-Learning or b-Learning spaces. In the latter, interaction features such as eye contact or comments about what a student or group of students may have done occurs in a much different way to what is found in face-to-face interaction. As a result, in most cases experienced teachers will need to have their digital skills updated. In contrast, inexperienced teachers lack this particular teaching background and normally possess more highly developed digital skills, such that their training should be geared more towards skills related to how to approach teaching in digital environments in terms of applying technologies they are already familiar with (Dumulescu et al., 2021; Kubik et al., 2021).

This study has also shown how interaction in online courses is complex because, although there is a synchronised phase and resources such as forums or chats are available, participant interaction is not always fluid. The same trend can also be seen in e-Learning or b-Learning teaching (Sáiz-Manzanares et al., 2017, 2021a,b). As a result, further research needs to be conducted into how fluent interaction can be improved, whether on the part of the teacher or the student, in order

TABLE 5 | Descriptive statistics and Mann–Whitney *U* test for the results in the satisfaction survey of the training activity in teachers participating in the satisfaction surveys for the synchronised sessions.

Final evaluation survey for the activity	Group 1 Inexperienced *n* = 8	Group 2 Experienced *n* = 13	*U* Mann–Whitney	*p*	*Z*	η^2
	M(*SD*)	*M*(*SD*)				
Training session 1. Definition and use of VLE						
1. The concepts dealt with in this section were clear to me.	4.67(0.47)	4.17(0.60)	29.00	0.08	−1.74	0.15
2. The materials presented in this session have proven useful for my teaching.	4.15(0.83)	4.25(0.54)	47.50	0.74	−0.34	0.01
3. The complementary information has proven to be useful to me.	4.42(0.50)	4.32(0.51)	51.50	0.97	−0.04	0.00
4. This session requires more work time.	3.59(1.50)	3.79(0.64)	51.50	0.97	−0.04	0.00
Training session 2. Design of materials and use of avatars						
1. The concepts dealt with in this section were clear to me.	4.82(0.37)	4.37(0.33)	18.00	0.009**	−2.62	0.34
2. The materials presented in this session have proven useful for my teaching.	4.68(0.71)	4.33(0.65)	27.50	0.06	−1.91	0.18
3. The complementary information has proven to be useful to me.	4.55(0.73)	4.28(0.50)	33.50	0.16	−1.41	0.10
4. This session requires more work time.	4.12(0.99)	3.81(0.68)	44.00	0.55	−0.60	0.02
Training session 3. Design of rubrics in VLE						
1. The concepts dealt with in this section were clear to me.	4.44(0.73)	4.54(0.32)	50.00	0.88	−0.15	0.00
2. The materials presented in this session have proven useful for my teaching.	4.32(1.17)	4.74(0.21)	47.50	0.72	−0.36	0.01
3. The complementary information has proven to be useful to me.	4.46(0.91)	4.80(0.17)	47.50	0.72	−0.36	0.01
4. This session requires more work time.	4.25(1.04)	3.85(0.99)	35.50	0.18	−1.34	0.09
Training session 4. Use of Learning Analytics Systems in VLE						
1. The concepts dealt with in this section were clear to me.	4.47(1.04)	4.30(0.27)	25.50	0.04*	−2.07	0.21
2. The materials presented in this session have proven useful for my teaching.	4.66(0.69)	4.62(0.32)	34.50	0.17	−1.39	0.10
3. The complementary information has proven to be useful to me.	4.66(0.69)	4.62(0.32)	34.50	0.17	−1.39	0.10
4. This session requires more work time.	4.42(0.49)	4.62(0.32)	48.00	0.76	−0.31	0.00
Training session 5. Dissemination in social networks						
1. The concepts dealt with in this section were clear to me.	4.81(0.20)	4.51(0.31)	24.00	0.02*	−2.35	0.28
2. The materials presented in this session have proven useful for my teaching.	4.88(0.13)	4.67(0.31)	30.00	0.06	−1.85	0.17
3. The complementary information has proven to be useful to me.	4.88(0.13)	4.67(0.31)	30.00	0.06	−1.85	0.17
4. This session requires more work time.	4.06(1.00)	3.76(0.88)	43.50	0.48	−0.71	0.03

*$p < 0.05$; **$p < 0.01$.
Z = Z de Kolmogorov–Smirnov; effect size η^2.

to offset the feeling of loneliness in the net. Achieving this might involve elements related to the use of social networks (de la Fuente et al., 2021b) as a resource for teaching.

In sum, there is still a long way to go before we achieve digital transformation in higher education in terms of the teaching-learning process. In order to accomplish this, further research is required exploring the acquisition of digital skills in university processes. One way of approaching this, in addition to updating teachers' digital skills, would involve including courses on digitalisation in the curricula of all university degrees. This is the challenge facing those responsible for higher education institutions the world over. In short, teaching staff need to be trained in how to design teaching activities that include SRL processes through ALT, since these resources allow the planning stage to be focused, for example with the use of avatars (Azevedo et al., 2015, Azevedo and Gašević, 2019), the follow-up stage (Azevedo and Gašević, 2019; Hosain et al., 2019; Noroozi et al., 2019), for example using tools similar to UBUMonitor (Sáiz-Manzanares et al., 2021b), and the self-evaluation stage (Kramarski and Michalsky, 2009; Bernardo et al., 2017), for example through the use of gamification activities so as to ultimately enhance motivation towards the goal of learning (Zimmerman, 2008). All of this will aid student development of metacognitive learning strategies when processing information (Carlson, 2003; Rothbart et al., 2011; McClelland et al., 2014; Valadas et al., 2017). This is one of the challenges facing teaching in the 21st century, since the mere use of LMS by no means ensures that deep-seated and reliable learning will be achieved (Yamada and Hirakawa, 2016; Park and Jo, 2017). As a result, training in digital skills, both in terms of their use and design, is the challenge facing educational institutions, particularly those engaged in higher education (Redecker and Punie, 2017; Jarillo et al., 2019; García-Peñalvo and Corell, 2020; García-Peñalvo, 2021).

Summing up, it can be concluded that further studies are needed that delve more deeply into a detailed analysis of instructional processes in digital skills based on self-regulation aimed at teaching staff. There are various resources, such as

TABLE 6 | Relation between the studies which served as the theoretical basis for the study and the outcomes to emerge from this work.

Previous studies	Results found in this study
Virtual learning environments help SRL (Azevedo et al., 2011, 2015) through various hypermedia resources, such as avatars and serious games. Studies by Kretschmer and Terharen (2019), Nappo et al. (2020), Sáiz-Manzanares et al. (2020), and Van De Weijer et al. (2020) found that the use of gamification enhances cognitive skills and boosts student motivation.	Prior to the training activity, both doctoral teaching staff and students alike displayed an interest in knowing the possible teaching resources that could be used in virtual contexts. They were also eager to know both the basic and the advanced techniques as well as the strategies that could be used in self-regulation. They also expressed a high degree of satisfaction at having taken part in the gamification activities.
Krouska et al. (2020) found that platforms such as Moodle were very useful and easy to use for teaching.	Participant satisfaction in this study with the resources applied in the virtual platform was high, both amongst doctoral teaching staff and students (future university teachers).
A key factor in the development of learning in LMS is to take particular care when devising resources and activities (Jommanop and Mekruksavanich, 2019; Kubik et al., 2021; Sáiz-Manzanares et al., 2021b). SRL is a key skill at all educational levels and can be boosted by developing structured training programmes aimed at teachers, particularly those who are undergoing their training (Dumulescu et al., 2021; Kubik et al., 2021).	Designing a training programme based on self-regulation has helped with the follow-up and analysis of the training process. This method has allowed for an analysis of the learning patterns developed by the participants, for which heat map visualisation techniques have been used. These resources help teachers to see differences in patterns easily and quickly.
The use of meta-tutoring or automatic tutoring resources in LMS aids student-centred learning, fosters student commitment and improves knowledge acquisition (Troussas et al., 2021; Krouska et al., 2021a).	This study reported a high degree of satisfaction with activities that included automatic feedback (e.g., gamification activities). Moreover, in this aspect no significant differences were found between doctoral teachers and students in terms of satisfaction regarding the use of these techniques.
Scheduled synchronous sessions boost collaborative work and assessment systems with feedback on the process (de la Fuente et al., 2021b).	The synchronous stages of the training process have been linked to greater participant access to the Moodle platform vs. less access in asynchronous sessions. Likewise, a difference was found in frequency of access between doctoral teaching staff and students; specifically, in favour of doctoral students with regard to content related to materials for designing rubrics in VLE and the use of Learning Analytics systems in VLE.
	Likewise, a difference was found in frequency of access between doctoral teaching staff and students; specifically, in favour of doctoral students with regard to content related to materials for designing rubrics in VLE and the use of Learning Analytics systems in VLE, although the motivation for the learning tasks was high in both group.
Teacher training in virtual environments is one of the goals of the 2030 Agenda (Redecker and Punie, 2017; Jarillo et al., 2019; García-Peñalvo, 2021) that has increased as a result of the current COVID-19 pandemic (Redecker and Punie, 2017).	This study offers materials and tools for designing training courses in digital skills for teaching based on self-regulated instruction in virtual environments.
It is important to gain an understanding of what both experienced and inexperienced teachers consider to be the strengths and weaknesses of LMS, which are key references in the current pandemic, particularly in higher education. The use of resources that include automatic personalised feedback procedures is key to enhancing student motivation. The interaction difference between synchronous and asynchronous sessions is evidenced (Collazos et al., in press).	This study offers a number of tools for gauging user perception of their satisfaction with learning processes in virtual environments. It also offers serious game materials for implementing automatic feedback on knowledge acquisition. Participant satisfaction with the training process has been evidenced (mean values of four out of five). Greater participation in synchronous than in asynchronous sessions has also been evidenced.

the use of serious games that include automatic feedback on the learning outcomes that can help with autonomous and personalised learning. Further work should also be carried out on developing resources that help to bridge the gap between participants' synchronous and asynchronous participation in educational activities. Teacher training, which is quite new in terms of these skills, will prove to be key if these teachers are to later use these tools in their everyday teaching practice.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Burgos Bioethical Committee, No. IR 27/2019, the coordinating university. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MS-M and LA: design and initial writing. LM-A and MS-M: review. MS-M, LA, MC and JV-B: final writing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded through the European "Self-Regulated Learning in SmartArt" Project 2019-1-ES01-KA204-065615.

ACKNOWLEDGMENTS

The authors would like to thank teachers at the universities of Burgos, Oviedo, Minho, and Valladolid for their participation in the Learning Activities within the SmartArt Project, as well as those teachers who took part as lecturers in the training activity. The authors are also grateful for the cooperation provided by the Centre for Virtual Teaching at the University of Burgos and for them having allowed the use of the UBUVirtual platform for the training activity. The authors also thank the teachers Yi Peng Ji, Raúl Marticorena-Sánchez, and Carlos Pardo Aguilar for developing the UBUMonitor tool.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.776806/full#supplementary-material

REFERENCES

Anguera, T. (1986). La investigacion cualitativa. *Educar* 10, 23–50.

Arthur, D., and Vassilvitskii, S. (2007). "K-means ++: The Advantages of Careful Seeding." in *Proceedings of the Eighteenth annual ACM-SIAM Symposium on Discrete Algorithms 2006*. January 22–26 2006; Miami, FL, USA, 1027–1035.

Atlas.ti. (2020). Software Package Qualitative Data Analysis; Version 8; Atlas.ti Scientific Software Development; GmbH: Berlin, Germany, 2020. Available at: https://atlasti.com/es/ (Accessed on 31 December 2020).

Azevedo, R., and Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: issues and challenges. *Comput. Hum. Behav.* 96, 207–210. doi: 10.1016/j.chb.2019.03.025

Azevedo, R., Johnson, A., Chauncey, A., and Graesser, A. (2011). "Use of hypermedia to assess and convey self-regulated learning," in *Handbook of Self-Regulation of Learning and Performance*. eds. B. J. Zimmerman and D. H. Schunk (New York: Routledge/Taylor & Francis Group), 102–121.

Azevedo, R., Taub, M., and Mudrick, N. (2015). "Technologies supporting self-regulated learning" in *The SAGE Encyclopedia of Educational Technology*. SAGE Thousand Oaks, CA, 731–734.

Bandalos, D. L., and Finney, S. J. (2001). "Item parceling issues in structural equation modeling," in *New Developments and Techniques in Structural Equation Modeling* eds. G. A. Marcoulides and R. E. Schumacker (New Yersey: Lawrence Erlbaum Associates Publishers), 269–296.

Bandura, A. (1986). Fearful expectations and avoidant actions as coeffects of perceived self-inefficacy. *Am. Psychol.* 41, 1389–1391. doi: 10.1037/0003-066X.41.12.1389

Bernardo, A., Cervero, A., Esteban, M., Tuero, E., Casanova, J. R., and Almeida, L. S. (2017). Freshmen program withdrawal: types and recommendations. *Front. Psychol.* 8:1544. doi: 10.3389/fpsyg.2017.01544

Boekaerts, M., and Niemivirta, M. (2000). "Self-regulated learning: finding a balance between learning goals and ego-protective goals," in *Handbook of Self-Regulation* eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (Academic Press), 417–450.

Calkins, S. D., and Williford, A. P. (2009). "Taming the terrible twos: self-regulation and school readiness," in *Handbook of Child Development and Early Education: Research to Practice*. eds. O. A. Barbarin and B. H. Wasik (New York: The Guilford Press), 172–198.

Campbell, D. F., and Stanley, J. C. (2005). *Diseños experimentales y cuasiexperimentales en la investigación social. [Experimental and quasi-experimental designs in social research. Experimental and quasi-experimental designs in social research]* Amorrortu editores.

Carbonero, M. A., Martín-Antón, L. J., Flores, V., and Freitas Resende, A. (2017). Estudio comparado de los estilos de enseñanza del profesorado universitario de ciencias sociales de España y Brasil. *Rev. Complut. de Educ.* 28, 631–647. doi: 10.5209/rev_RCED.2017.v28.n2.50711

Carlson, S. M. (2003). Executive function in context: development, measurement, theory, and experience. *Monogr. Soc. Res. Child Dev.* 68, 138–151. doi: 10.1111/j.1540-5834.2003.06803012.x

Castañer-Balcells, M., Camerino-Foguet, O., and Anguera-Argilaga, M. T. (2013). Métodos mixtos en la investigación de las ciencias de la actividad física y el deporte. *Apunts Educación Física y Deportes* 112, 31–36. doi: 10.5672/apunts.2014-0983.es.(2013/2).112.01

Cervone, D. (1993). "The role of self-referent cognitions in goal setting, motivation, and performance" in *Cognitive Science Foundations of Instruction* ed. M. Rabinowitz (New York: Lawrence Erlbaum Associates, Inc), 57–95.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* New York: Routledge.

Collazos, C. A., Fardoun, H., AlSekait, D., Santos Pereira, C., and Moreira, F. (2021). Designing online platforms supporting emotions and awareness. *Electronics* 10:251. doi: 10.3390/electronics10030251

Collazos, C. A., Pozzi, F., and Romagnoli, M. (in press). The use of e-learning platforms in a lockdown scenario – a study in Latin American countries. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 16, 419–423. doi: 10.1109/RITA.2021.3137632

de la Fuente, J., Kauffman, D. F., Dempsy, M. S., and Kauffman, Y. (2021a). Analysis and Psychoeducational implications of the behavior factor During the COVID-19 emergency. *Front. Psychol.* 12:613881. doi: 10.3389/fpsyg.2021.613881

de la Fuente, J., Pachón-Basallo, M., Santos, F. H., Peralta-Sánchez, F. J., González-Torres, M. C., Artuch-Garde, R., et al. (2021b). How has the COVID-19 crisis affected the academic stress of university students? The role of teachers and students. *Front. Psychol.* 12:626340. doi: 10.3389/fpsyg.2021.626340

Dobashi, K., Ho, C. P., Fulford, C. P., and Lin, M. F. G. (2019). "A heat map generation to visualize engagement in classes using Moodle learning logs." in *Proceedings of 2019 4th International Conference on Information Technology: Encompassing Intelligent Technology and Innovation Towards the New Era of Human Life, InCIT 2019*, 138–143.

Dumulescu, D., Pop-Pă Curar, I., and Constantin Valer Necula, C. (2021). Learning Design for Future Higher Education – insights from the time of COVID-19. *Front. Psychol.* 12:647948. doi: 10.3389/fpsyg.2021.647948

Enembreck, F., and Barthès, J.-P. (2005). "Personalization in multi-agent systems." in *Proceedings - 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, IAT'05, 2005, 230–233.

Flick, U. (2014). *El diseño de la investigación cualitativa* Madrid: Ediciones Morata.

García-Peñalvo, F. J. (2021). Digital transformation in the universities: implications of the COVID-19 pandemic. *Educ. Knowl. Soc* 22, 1–6. doi: 10.14201/eks.25465

García-Peñalvo, F. J., and Corell, A. (2020). La COVID-19: ¿enzima de la transformación digital de la docencia o refejo de una crisis metodológica y competencial en la educación superior? [The CoVId-19: the enzyme of the digital transformation of teaching or the reflection of a methodological and competence crisis in higher education?]. *Campus Virtuales*, 9, 83–98.

Hosain, A. A., Santhalingam, P. S., Pathak, P., Kosecka, J., and Rangwala, H. (2019). "Sign language recognition analysis using multimodal data." in *Proceedings – 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019*, October 5-8, 2019. 203–210.

Huizinga, M., Baeyens, D., and Burack, J. A. (2018). Editorial: executive function and education. *Front. Psychol.* 9:1357. doi: 10.3389/fpsyg.2018.01357

Hull, A., Boulay, B., and du Boulay, B. (2015). Motivational and metacognitive feedback in SQL-tutor*. *Comput. Sci. Educ.* 25, 238–256. doi: 10.1080/08993408.2015.1033143

IBM (2016). SPSS Statistical Package for the Social Sciences (SPSS); Version 24. Madrid: Spain.

Jarillo, M. P., Pedraza, L., Ger, P. M., and Bocos, E. (2019). Challenges of online higher education in the face of the sustainability objectives of the united nations: carbon footprint, accessibility and social inclusion. *Sustainability* 11. doi: 10.3390/su11205580

Ji, Y. P., Marticorena-Sánchez, R., and Pardo-Aguilar, C. (2018). UBUMonitor: Monitoring of students on the Moodle platform. Available at: https://github.com/yjx0003/UBUMonitor (Accessed 24 December 2020).

Jommanop, T., and Mekruksavanich, S. (2019). "E-learning recommendation model based on multiple intelligence." in *Proceedings – 2019 14th international joint symposium on artificial intelligence and natural language processing, ISAI-NLP 2019*.

Kramarski, B., and Michalsky, T. (2009). Investigating preservice teachers' professional growth in self-regulated learning environments. *J. Educ. Psychol.* 101, 161–175. doi: 10.1037/a0013101

Kretschmer, V., and Terharen, A. (2019). "Serious games in virtual environments: Cognitive ergonomic trainings for workplaces in intralogistics," in *Advances in Human Factors in Wearable Technologies and Game Design*. ed. T. Z. Ahram (Cham: Springer International Publishing), 266–274.

Krouska, A., Troussas, C., Giannakas, F., and Sgouropoulou, C. (2021a). "Enhancing the effectiveness of intelligent tutoring systems using adaptation and cognitive diagnosis modeling," in *Novelties in Intelligent Digital Systems*. eds. C. Frasson et al. (Amsterdam, Berlin, Washington, DC: IOS Press Ebooks), 40–45.

Krouska, A., Troussas, Ch., and Sgouropoulou, C. (2020). "Usability and Educational Affordance ofWeb 2.0 Tools from Teachers' Perspectives." In *PCI 2020: 24th Pan-Hellenic Conference on Informatics Athens Greece*. eds. N.N. Karanikolas and A. Voulodimos. 20–22 November 2020. Association for Computing Machinery.

Krouska, A., Troussas, C., and Sgouropoulou, C. (2021b). A cognitive diagnostic module based on the repair theory for a personalized user experience in E-learning software. *Computers* 10, 1–12. doi: 10.3390/computers10110140

Kubik, V., Frey, I.-G., and Gaschler, R. (2021). PLAT 20(3) 2021: Promoting self-regulated learning: training, feedback, and addressing teachers' misconceptions. *Psychol. Learn. Teach.* 20, 306–323. doi: 10.1177/14757257211036566

Liew, J. (2012). Effortful control, executive functions, and education: bringing self-regulatory and social-emotional competencies to the table. *Child Dev. Perspect.* 6, 105–111. doi: 10.1111/j.1750-8606.2011.00196.x

López-Roldán, P, and Fachelli, S. (2015). Metodología de la Investigación Social Cuantitativa. Universidad Autónoma de Barcelona.

Lumsden, J., Skinner, A., Coyle, D., Lawrence, N., and Munafo, M. (2017). Attrition from web-based cognitive testing: A repeated measures comparison of gamification techniques. *J. Med. Internet Res.* 19, e395–e320. doi: 10.2196/jmir.8473

Mann, H. B., and Whitney, D. R. (1947). On a test of Whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

Martín-Antón, L. J., Carbonero, M. A., Valdivieso, J. A., and Monsalvo, E. (2020). Influence of Some personal and family variables on social responsibility Among primary education students. *Front. Psychol.* 11:1124. doi: 10.3389/fpsyg.2020.01124

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., et al. (2014). Predictors of early growth in academic achievement: the head-toes-knees-shoulders task. *Front. Psychol.* 5:599. doi: 10.3389/fpsyg.2014.00599

McClelland, M. M., Cameron, C. E., and Tominey, S. L. (2010). "elf-regulation : The integration of cognition and emotion," in *Handbook of Life-Span Development*. eds. R. M. Lerner and W. Overton (Wiley & Sons), 509–553.

Nappo, R., Iorio, M., and Somma, F. (2020). "STRAS: A Software for the Assessment and Training of Executive Functions in Children." in *Conference: Proceedings of the First Symposium on Psychology-Based Technologies at Naples*", September 25-26, 2019.

Noroozi, O., Järvelä, S., and Kirschner, P. A. (2019). Multidisciplinary innovations and technologies for facilitation of self-regulated learning. *Comput. Hum. Behav.* 100, 295–297. doi: 10.1016/j.chb.2019.07.020

Nurmi, J., Knittle, K., Ginchev, T., Khattak, F., Helf, C., Zwickl, P., et al. (2020). Engaging users in the behavior change process with digitalized motivational interviewing and gamification: development and feasibility testing of the precious app. *JMIR Mhealth Uhealth* 8:e12884. doi: 10.2196/12884

Páramo Fernández, M. F., Araújo, A. M., Tinajero Vacas, C., Almeida, L. S., and Rodríguez González, M. S. (2017). Predictors of students' adjustment during transition to university in Spain. *Psicothema* 29, 67–72. doi: 10.7334/psicothema2016.40

Park, Y., and Jo, I. H. (2017). Using log variables in a learning management system to evaluate learning activity using the lens of activity theory. *Assess. Eval. High. Educ.* 42, 531–547. doi: 10.1080/02602938.2016.1158236

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., and Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemp. Educ. Psychol.* 36, 36–48. doi: 10.1016/j.cedpsych.2010.10.002

Pinnell, C. (2015). *Computer Games for Learning: an Evidence-Based Approach*, vol. 18, 523–524.

Raftery, J. N., and Bizer, G. Y. (2009). Negative feedback and performance: The moderating effect of emotion regulation. *Personal. Individ. Differ.* 47, 481–486. doi: 10.1016/j.paid.2009.04.024

Redecker, C., and Punie, Y. (2017). *European Framework for the Digital Competence of Educators: DigCompEdu*; European Union: Brussels, Belgium; Luxembourg Publications

Remesal, A., Colomina, R. M., Mauri, T., and Rochera, M. J. (2017). Online questionnaires use with automatic feedback for e-innovation in university students / Uso de cuestionarios online con feedback automático para la e-innovación en el alumnado universitario. *Comunicar* 25, 51–60. doi: 10.3916/C51-2017-05

Rheinberg, F., Vollmeyer, R., and Rollett, W. (2000). "Chapter 15 - motivation and action in self-regulated learning," in *Handbook of Self-Regulation*. eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (New York, London: Academic Press), 503–529.

Rothbart, M. K., Sheese, B. E., Rueda, M. R., and Posner, M. I. (2011). Developing mechanisms of self-regulation in early life. *Emot. Rev.* 3, 207–213. doi: 10.1177/1754073910387943

Sáiz-Manzanares, M. C. (2021). Ad hoc survey was drawn up to ascertain participants' level of prior knowledge of the training activity related to their know-how and application of teaching resources in virtual learning environments. Unpublished document.

Sáiz-Manzanares, M. C., García Osorio, C. I., Díez-Pastor, J. F., and Martín Antón, L. J. (2019b). Will personalized e-learning increase deep learning in higher education? *Inf. Discov. Deliv.* 47, 53–63. doi: 10.1108/IDD-08-2018-0039

Sáiz-Manzanares, M. C., García-Osorio, C. I., and Díez-Pastor, J. F. (2019a). Differential efficacy of the resources used in B-learning environments. *Psicothema* 31, 170–178. doi: 10.7334/psicothema2018.330

Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., García Osorio, C. I., and Díez-Pastor, J. F. (2017). How do B-learning and learning patterns influence learning outcomes? *Front. Psychol.* 8:745. doi: 10.3389/fpsyg.2017.00745

Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., Muñoz-Rujas, N., Rodríguez-Arribas, S., Escolar-Llamazares, M. C., Alonso-Santander, N., et al. (2021a). Teaching and learning styles on Moodle: An analysis of the effectiveness of using stem and non-stem qualifications from a gender perspective. *Sustainability* 13, 1–21. doi: 10.3390/su13031166

Sáiz-Manzanares, M. C., Rodríguez-Arribas, S. R., Pardo-Aguilar, C., and Queiruga-Dios, M. Á. (2020). Effectiveness of self-regulation and serious games for learning stem knowledge in primary education. *Psicothema* 32, 516–524. doi: 10.7334/psicothema2020.30

Sáiz-Manzanares, M. C., Rodríguez-Díez, J. J., Díez-Pastor, J. F., Rodríguez-Arribas, S., Marticorena-Sánchez, R., and Ji, Y. P. (2021b). Monitoring of student learning in learning management systems: An application of educational data mining techniques. *Appl. Sci.* 11, 1–16. doi: 10.3390/app11062677

Schunk, D. H. (1995). "Self-efficacy and education and instruction," in *Self-Efficacy, Adaptation, and Adjustment. The Plenum Series in Social/Clinical*

Psychology; Naples, September 25–26, 2019. ed. J. E. Maddux (Boston, MA: Springer), 281–303.

Schunk, D. H. (1996). Goal and self-evaluative influences During Children's cognitive skill learning. *Am. Educ. Res. J.* 33, 359–382. doi: 10.3102/00028312033002359

Schunk, D. H., and Ertmer, P. A. (2000). "Chapter 19: Self-regulation and academic learning: self-efficacy enhancing interventions," in *Handbook of Self-Regulation*. eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo: Academic Press), 631–649.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., and Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learn. Instr.* 54, 93–103. doi: 10.1016/j.learninstruc.2017.08.005

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., and Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with CRYSTAL ISLAND. *Comput. Hum. Behav.* 76, 641–655. doi: 10.1016/j.chb.2017.01.038

Troussas, C., Krouska, A., and Sgouropoulou, C. (2021). A novel teaching strategy Through adaptive learning activities for computer programming. *IEEE Trans. Educ.* 64, 103–109. doi: 10.1109/TE.2020.3012744

Valadas, S. T., Almeida, L. S., and Araújo, A. M. (2017). The mediating effects of approaches to learning on the academic success of first-year college students. *Scand. J. Educ. Res.* 61, 721–734. doi: 10.1080/00313831.2016.1188146

Van De Weijer, S. C. F., Duits, A. A., Bloem, B. R., De Vries, N. M., Kessels, R. P. C., Köhler, S., et al. (2020). Feasibility of a cognitive training game in Parkinson's disease: the randomized Parkin'Play study. *Eur. Neurol.* 83, 426–432. doi: 10.1159/000509685

Weinstein, C. E., and Acee, T. W. (2018). "Study and learning strategies," in *Handbook of College Reading and Study Strategy Research*. eds. I. R. F. Flippo and T. W. Bean (New York: Routledge), 227–240.

Yamada, Y., and Hirakawa, M. (2016). "A case study of analyzing logs of LMS in flipped classroom." in *Proceedings – 2015 IIAI 4th international congress on advanced applied informatics, IIAI-AAI 2015*, 374–378.

Zeynali, S., Pishghadam, R., and Hosseini Fatemi, A. (2019). Identifying the motivational and demotivational factors influencing students' academic achievements in language education. *Learn. Motiv.* 68:101598. doi: 10.1016/j.lmot.2019.101598

Zimmerman, B. J. (2000). *Attaining Self-Regulation: A Social Cognitive Perspective. In Handbook of Self-Regulation* San Diego: Academic Press, 13–39.

Zimmerman, B. J. (2005). Can CBLEs be used as SRL tools to enhance learning? *Educ. Psychol.* 40, 199–209. doi: 10.1207/s15326985ep4004

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* 45, 166–183. doi: 10.3102/0002831207312909

Zimmerman, B. J. (2013). From cognitive modeling to self-Regul+ation: A social cognitive career path. *Educ. Psychol.* 48, 135–147. doi: 10.1080/00461520.2013.794676

Zimmerman, B. J., and Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance*. Routledge/Taylor & Francis Group.

Zorrilla Pantaleón, M. E., García-Saiz, D., and de la Vega, A. (2021). Fostering study time outside class using gamification strategies: An experimental study at tertiary-level database courses. *Comput. Appl. Eng. Educ.* 29, 1340–1357. doi: 10.1002/cae.22389

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 23 May 2022
doi: 10.3389/feduc.2022.818365

Performance and Configuration of Artificial Intelligence in Educational Settings. Introducing a New Reliability Concept Based on Content Analysis

Florian Berding[1], Elisabeth Riebenbauer[2], Simone Stütz[3], Heike Jahncke[4], Andreas Slopinski[4] and Karin Rebmann[4]*

[1] Department of Professional Education and Life-Long Learning, Faculty of Education, University of Hamburg, Hamburg, Germany, [2] Department of Business Education and Development, School of Business, Economics and Social Sciences, University of Graz, Graz, Austria, [3] Institute for Business and Vocational Education and Training, Johannes Kepler University Linz, Linz, Austria, [4] Business Administration and Business Education, Department of Business Administration, Economics and Law, University of Oldenburg, Oldenburg, Germany

OPEN ACCESS

Edited by:
Manuel Gentile, Istituto per le Tecnologie Didattiche (CNR-ITD), Italy

Reviewed by:
Josef Guggemos, University of St. Gallen, Switzerland
Jose Ramon Saura, Rey Juan Carlos University, Spain

***Correspondence:**
Florian Berding
florian.berding@uni-hamburg.de

Specialty section:
This article was submitted to Digital Learning Innovations, a section of the journal Frontiers in Education

Received: *19 November 2021*
Accepted: *02 May 2022*
Published: *23 May 2022*

Citation:
Berding F, Riebenbauer E, Stütz S, Jahncke H, Slopinski A and Rebmann K (2022) Performance and Configuration of Artificial Intelligence in Educational Settings. Introducing a New Reliability Concept Based on Content Analysis. Front. Educ. 7:818365. doi: 10.3389/feduc.2022.818365

Learning analytics represent a promising approach for fostering personalized learning processes. Most applications of this technology currently do not use textual data for providing information on learning, or for deriving recommendations for further development. This paper presents the results of three studies aiming to make textual information usable. In the first study, the iota concept is introduced as a new content analysis measure to evaluate inter-coder reliability. The main advantage of this new concept is that it provides a reliability estimation for every single category, allowing deeper insight into the quality of textual analysis. The second study simulates the process of content analysis, comparing the new iota concept with well-established measures (e.g., Krippendorff's Alpha, percentage agreement). The results show that the new concept covers the true reliability of a coding scheme, and is not affected by the number of coders or categories, the sample size, or the distribution of data. Furthermore, cut-off values are derived for judging the quality of the analysis. The third study employs the new concept, as it analyzes the performance of different artificial intelligence (AI) approaches for interpreting textual data based on 90 different constructs. The texts used here were either created by apprentices, students, and pupils, or were taken from vocational textbooks. The paper shows that AI can reliably interpret textual information for learning purposes, and also provides recommendations for optimal AI configuration.

Keywords: learning analytics, artificial intelligence, content analysis, reliability, hyperparameter, neural net, decision trees, random forest

INTRODUCTION[1]

Meta- and meta-meta analyses show that the integration of digital technologies increases the efficiency and effectiveness of learning processes (Kulik and Kulik, 1991; Means et al., 2010; Tamim et al., 2011; Bernard et al., 2014). Several meta-analyses have proven the usefulness of design principles for multimedia learning environments (Brom et al., 2018; Schneider et al., 2018; Mayer, 2019; Mayer and Fiorella, 2019; Mayer and Pilegard, 2019; Alpizar et al., 2020), and digital technologies are critical for designing state-of-the-art instructional processes.

The improvement potential offered by digital technologies can be enhanced even further if the design of instruction is adapted to the individual prerequisites of every single learner. The advantages of personalized instruction have been empirically supported by several studies (Schrader, 1989; Anders et al., 2010; Karst et al., 2014). For example, Bloom (1984) showed that individual tutoring is more effective than traditional classroom settings with 30 students per teacher. A study by VanLehn (2011) shows that computer-based intelligent tutoring systems are nearly as effective as one-on-one human tutoring.

One possibility for implementing personalized learning is via *learning analytics* which aims to improve learning (Rienties et al., 2020). These are "the collection, analysis, and application of data accumulated to assess the behavior of educational communities. Whether it be through the use of statistical techniques and predictive modeling, interactive visualizations, or taxonomies and frameworks, the ultimate goal is to optimize both student and faculty performance, to refine pedagogical strategies, to streamline institutional costs, to determine students' engagement with the course material, to highlight potentially struggling students (and to alter pedagogy accordingly), to fine tune grading systems using real-time analysis, and to allow instructors to judge their own educational efficacy" (Larusson and White, 2014). The actual practice of learning analytics was reported in a literature review of 401 research papers by Jaakonmäki et al. (2020), showing that they are mostly applied for the evaluation of student performance, decision support, and clustering of learners. However, it was determined that the real-time analysis of students' learning behavior, and the adaption of learning materials and demands to individual needs are only rarely conducted.

The reason for this low level of personalization can be traced to the high organizational and technical demands of implementation. This type of learning analytics represents the second-to-last level of organizational implementation in the learning analytics sophistication model proposed by Siemens et al. (2013). Another reason is the limited quality of data available for the purpose of learning analytics. For example, many studies use so-called log data, which represents the interaction of a learner with the learning environment. This includes elements such as the number of assessment attempts, time taken for assessments, videos seen, or videos viewed repeatedly (Ifenthaler and Widanapathirana, 2014; Liu et al., 2018; ElSayed et al., 2019). Other studies opt for a research approach to learning analytics that is based on the analysis of stable and/or historical data such as students' social backgrounds and demographic characteristics, historical education records, or average historical grades (Ifenthaler and Widanapathirana, 2014; ElSayed et al., 2019). In their literature review, ElSayed et al. (2019) reported four additional data types that are used less frequently: multimodal data (e.g., heart rate, eye tracking), chat and forum conversations, video recordings, and self-reported data (e.g., questionnaires, interviews). On the one hand these data types are important for understanding individual learning, as well as for providing recommendations for further development, because empirical studies prove their predictive power. On the other hand this kind of data only provides limited insights about changes in students' cognition and motivation as the analysis of the students' interactions in terms of clicking in a digital learning environment does not provide enough ground for pedagogical decision-making (Reich, 2015).

What can be concluded from these studies is that data should be supplemented by textual data allowing a deeper analysis of the *quality* of learning processes and their outcomes. It is not only important to gather information on grades, gender, or how often a student repeats a video. It is also essential for fine-tuning future learning processes to understand which individual abilities, attitudes, and beliefs lead to current learning behavior and outcomes. Textual data can provide this kind of insight. For example, if teachers want to clarify whether their students have the "correct" understanding of "price" in an economy context, they could ask the students to write an essay in which they explain what a price is. The teachers can use this information to find a starting point for further instruction, especially if some students understand the concept in a "wrong" manner. Another example of this idea can be found in teacher education. Prospective teachers create learning materials containing textual data, such as learning task, explanations, and visualizations for a lesson plan. The information included in the textual components here strongly predicts what kind of learning processes a prospective teacher intends to apply. For example, the task "What kind of product assortment expansion can be characterized as 'diversification'?" does not include any of the experiences of apprentices, i.e., it is a de-contextual task. In contrast, the task "Explain the factors that influence the range of goods in your training company and discuss it with your colleagues" explicitly refers to the experience apprentices gain at the company where they are doing their training. Based on the textual information of the task, a teacher educator can conclude the extent in which prospective teachers integrate the experiences of their learners when creating a learning environment, and further interventions can be planned based on their conclusions.

Intervention planning makes it necessary to sort information into pedagogical and didactical theories. As Wong et al. (2019) state: "(. . .) [L]earning analytics require theories and principles on instructional design to guide the transformation of the information obtained from the data into useful knowledge for instructional design" (see also Luan et al., 2020). This complex challenge is illustrated in **Figure 1**. With learning analytics applications, the computer program has to understand the textual information, summarize the information in categories

[1] A preprint of this manuscript was published 03/2022 as Berding et al. (2022).

of scientific models and theories, and derive the impact of the categories on further learning to provide recommendations for learners and teachers. In essence, learning analytics applications have to solve the same problems as human teachers: diagnose the preconditions of learners, and tailor adequately adapted learning processes based on scientific insights.

Learning analytics require the realization of complex tasks using artificial intelligence (AI). AI describes the attempt to simulate human actions by a computer (Kleesiek et al., 2020), and consists of machine learning (ML). In ML, a computer solves a problem by developing the necessary algorithm itself (Alpaydin, 2019; Lanquillon, 2019). With the different types of ML, supervised machine learning is able to realize the model of **Figure 1**, providing links to established scientific models and theories. In this special case, AI attempts to generate a prediction model which transforms input data into output data. In the model seen in **Figure 1**, the first step aims to sort the information of an individual learner based on textual data into models and theories. The input data represents texts (e.g., written essays, interviews, tasks, instructional texts), while the output data represents categories from didactical and pedagogical theories and models (AI I). The next step predicts further learning and outcomes based on the identified categories (AI II). In this case, the input data are the categories, and the output data are characteristics of other learning-related variables (e.g., grades, motivation, use of learning strategies). Finally, the information about the learning-related variables forms the input data for generating recommendations as output data (AI III). In this stage, AI can recommend interventions that produce the strongest impact for the variable relevant for learning based on the current state of these variables. For example, if a student has low grades and low motivation, AI can recommend interventions that promote the quality and quantity of motivation based on the self-determination theory of motivation (Ryan and Deci, 2012), such as an informative feedback or granting students freedom while working on a task (Euler and Hahn, 2014). The increased motivation increases the chance that the students improve their grades since motivation is related to the quality of actions (Cerasoli et al., 2014).

This paper focuses on the first step of this process (AI I). AI has to understand textual data and learn whether and how this information belongs to scientific categories. AI here requires a data collection of input *and* output data for identifying the relationship between the two data types (Lanquillon, 2019). AI essentially has to conduct parts of a content analysis by assigning texts (input data) to categories (output data) based on an initial content analysis of humans. As this paper concentrates on supervised machine learning, this means that humans have to develop a coding scheme. That is, humans have to define the categories to which the text can be assigned. They have to ensure sufficient quality of the coding scheme, and they need to have applied the coding scheme to a specific number of textual documents in order to generate the necessary input and output data for the training of AI. Only on the basis of this data, AI can learn to conduct a content analysis which is limited to the coding processes of a human developed coding scheme. As a result, the quality of the training data for AI is critical as Song et al. (2020) recognized in their simulation study. In their study, the quality of the initial data accounts for 62% of the variance of the mean absolute prediction error.

Because the quality of content analysis performed by humans and computers is critical for the process of learning analytics, the accuracy of the assignments has to be very high, meaning a powerful AI algorithm that includes a configuration that optimizes its accuracy has to be selected. This also requires an accurate initial content analysis by humans. Whereas a large number of studies compare the performance of different kinds of AI (e.g., Lorena et al., 2011; Hartmann et al., 2019), different configurations of parameters have rarely been investigated (e.g., Probst et al., 2019). These hyperparameters have to be chosen before the learning process of AI begins; they are normally not optimized during the learning process (Probst et al., 2019). Furthermore, most performance studies do not analyze how accurately AI interprets the texts of students for learning purposes. Previous studies analyze textual data such as product reviews on Amazon, social media comments on Facebook or user generated content on Twitter (Hartmann et al., 2019; Saura et al., 2022). As a consequence, there is a clear research gap as there is no empirical evidence how well AI can be used for the analysis of textual data generated in educational settings.

The issue of determining the performance of AI for interpreting texts generally increases, because there is no widely-accepted performance measure for content analysis reliability regardless whether it is conducted by human or artificial intelligence. Reliability is a central characteristic of any assessment instrument, and describes the extent to which the instrument produces error-free data (Schreier, 2012). Krippendorff (2019) suggests replicability as a fundamental reliability concept, which is also referred to as *inter-coder reliability*. This describes the degree to which "a process can be reproduced by different analysts, working under varying conditions, at different locations, or using different but functionally equivalent measuring instruments" (Krippendorff, 2019). Past decades have seen a large number of reliability measures being suggested. The study by Hove et al. (2018) shows that the 20 reliability measures they investigated differ in their numeric values for the same data. Thus, it is hard to decide which measure to trust for the judgment of quality in content analysis. Krippendorff's Alpha is currently the most recommended reliability measure (Hayes and Krippendorff, 2007), as it can be applied to variables of any kind (nominal, ordinal, and metric); to any number of coders; to data with missing cases and unequal sample sizes; all while comprising chance correction (Krippendorff, 2019). Recent years, however, have seen the advantages of Krippendorff's Alpha being questioned and controversially discussed (Feng and Zhao, 2016; Krippendorff, 2016; Zhao et al., 2018). Zhao et al. (2013) analyzed different reliability measures, concluding that Krippendorff's Alpha contains problematic assumptions and produces the highest number of paradoxes and abnormalities. For example, they argue that Alpha penalizes improved coding, meaning that if coders correct errors, the values for Alpha can decrease (Zhao et al., 2013). Furthermore, cases exist where coder agreement is nearly 100%, while the Alpha values are about 0, indicating the

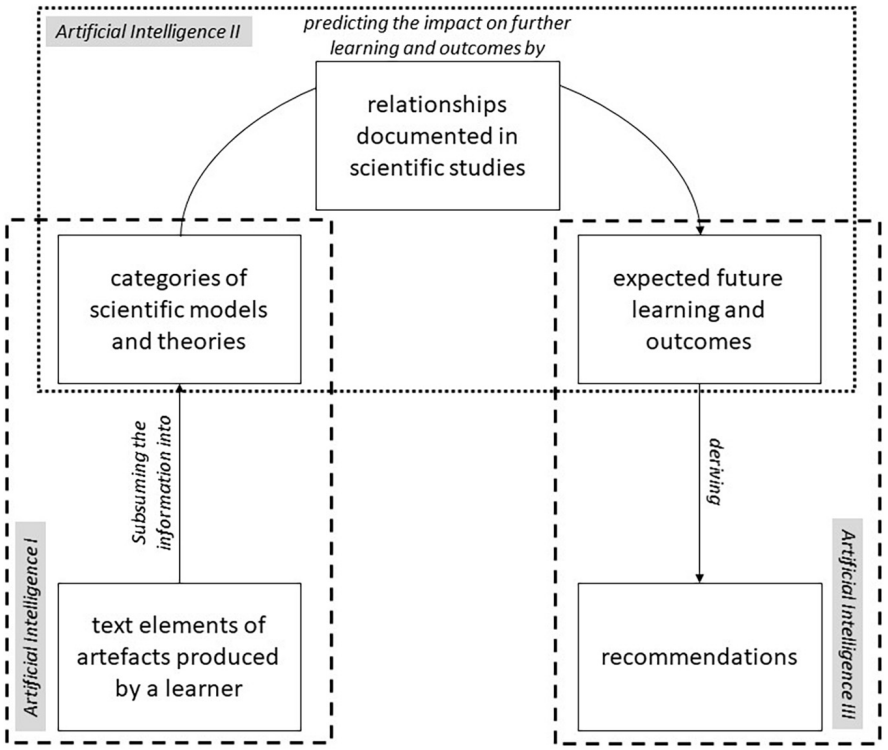


FIGURE 1 | Summarizing information on individual learners into models and theories for deriving recommendations.

absence of reliability. Thus, Krippendorff's Alpha may lead to false conclusions about the reliability of a content analysis. This is problematic since this measure has become one of the most used measures in content analysis in the last 30 years (Lovejoy et al., 2016) and is used in simulation studies for estimating the initial data's impact on the performance quality of AI (Song et al., 2020). As a result, there is a need for new reliability measures that overcome these difficulties (Zhao et al., 2013).

Feng and Zhao (2016) suggest to orientate a new reliability measure on the item response theory and not on the classical test theory. In the classical test theory reliability is characterized with measures such as Cronbach's Alpha. These measures produce a single numeric value for a complete scale similar to the measure currently used in content analysis (e.g., Krippendorff's Alpha, percentage agreement, Scott's Pi, Cohen's Kappa) (Lovejoy et al., 2016). From the perspective of the item response theory, this is an oversimplification since the reliability is not constant over the range of a scale. With the help of the test information curve, the reliability of a scale can be investigated for different scale characteristics (e.g., de Ayala, 2009; Baker and Kim, 2017). For example, a test for measuring the motivation of students can be more reliable in the middle than for the extreme poles implying that the test is reliable only for participants with medium motivation and less reliable for students with very low or very high motivation. Furthermore, some models of the item response theory such as Rasch models offer the opportunity to investigate if a scale produces a bias for different groups of individuals. That is, they allow to examine whether an instrument functions similarly for different groups of people (subgroup invariance) or not (e.g., Baker and Kim, 2017). Based on the previous example in this paragraph, a test may be more reliable for women with high motivation than for men with high motivation, leading to bias. Men with a high motivation may be falsely represented in the data. Current measures for content analysis do not provide these analytical opportunities.

In this context this paper has the following objectives:

(1) Developing a new performance measure for content analysis,
(2) Investigating and comparing the properties of the new measure with well-established measures,
(3) Analyzing the performance of AI based on the new measure, and deriving insights for the optimized configuration of AI in educational contexts.

By working on these objectives the originality of the present study is that

- it develops a new and innovative measure for content analysis based on the ideas of item response theory. That is, a measure that allows to assess the reliability of *every single* category of a coding scheme. Previous measures are limited to the scale level only.
- it develops a new measure for content analysis avoiding the problematic assumptions Krippendorff's Alpha uses as discussed in literature (Zhao et al., 2013, 2018; Feng and Zhao, 2016; Krippendorff, 2016).

- it generates rules of thumb for the new measure to judge the quality of content analysis in practical applications.
- it applies a new and innovative approach for determining the performance of AI in the interpretation of textual data produced within educational settings.

Thus, this paper aims to contribute to a progression in the field of content analysis by transferring the basic ideas of the item response theory to content analysis and by offering an additional tool for understanding how AI generates new information based on textual data.

In order to reach these objectives, section "Development of the New Inter-coder Reliability Concept" presents the mathematical derivation of the new concept called Iota Reliability Concept. In order to prove if the new concept is really a progression, section "Simulation Study of the New Reliability Concept" presents a simulation study simulating 808,500 coding tasks with a varying number of coders and categories and varying sample sizes. With the help of the simulation, the new measure is compared with percentage agreement which represents the most intuitive measure of inter-coder-reliability, and with Krippendorff's Alpha which represents the current state of research (Hayes and Krippendorff, 2007; Lovejoy et al., 2016). The simulation is also used to derive rules of thumb for judging the quality of content analysis in practical applications.

Section "Analyzing the Performance and Configuration of Artificial Intelligence" applies both the new and the established measures to real world cases by training three different types of AI to interpret 90 different didactical constructs. The data comprises essays written by students of different degrees and textual material out of textbooks. Training AI utilities *mlr3* (Lang et al., 2019) which is the newest framework for machine learning in the statistical coding language *R*. This provides insights into the performance of AI for educational purposes.

The paper ends with a discussion of the results and provides recommendations for researchers and practitioners. Section "Conclusion" provides an example for the analysis of AI with the new measures in order to demonstrate the potentials of the new concept.

DEVELOPMENT OF THE NEW INTER-CODER RELIABILITY CONCEPT

Overview

The aim of this new concept is to develop a reliability measure that provides information on every single category. To achieve this goal, we suggest a reliability concept consisting of three elements for *every* category: the alpha-, beta-, and iota-elements. The concept additionally provides an assignment-error matrix (AEM) offering information on how errors in the different categories influence the data in the others.

Reliability describes the extent of the absence of errors (Schreier, 2012), meaning the basic idea behind the alpha and beta elements is to take two different types of errors into account. These are described from the perspective of every single category. The alpha elements refer to the error of a coding unit being unintentionally assigned to the wrong category, e.g., when a unit is not assigned to A, although it belongs to A. The beta elements consider the error that a coding unit belonging to another category is unintentionally assigned to the category under investigation, e.g., when a unit is assigned to A, although it does not belong there.

This concept is based on six central assumptions:

(1) The core of content analysis is a scheme guiding coders to assign a coding unit to a category. Here, reliability is a property of a coding scheme, not of coders.
(2) The categories form a nominal or ordinal scale with discrete values.
(3) Every coding unit can be assigned to exactly one category.
(4) Every coding unit is assignable to at least one category.
(5) Coders judge the category of a coding unit by using a coding scheme or by guessing.
(6) Reliability can vary for each category.

The following sections systematically introduce the new concept and each of its elements.

Alpha Elements: Alpha Reliability and Alpha Error

Developing a reliability concept that reflects the reliability of the coding scheme for each single category requires the focus to be shifted from all data to the data that involves the category under investigation. **Figure 2** illustrates this idea for the case of two coders and three categories.

The gray cells in the tables show the relevant combinations for the categories. For example, in the table on the left, only the first row and the first column comprise coding judgments that involve category one. In the middle table, the gray cross represents all relevant coding for category two. The third row and the third column in the right table include coding for category three. The diagonal of the table shows all judgments for a category that the two coders agree on. For example, both coders agree that 120 coding units belong to category one, that five units belong to category two, and three coding units belong to category three.

The *alpha reliability* and the *alpha error* can be introduced based on this data and category perspective. The alpha reliability uses two basic ideas. First, the number of coding units all coders agree on for a specific category (e.g., 120 for category one, 5 for category two, etc.) represents the agreement of the coders regarding that category. Second, the number of all coding units that involve the specific category (e.g., $12 + 10 + 120 + 78 + 7$ for category one) is an approximation of the number of coding units that belong to the specific category. Thus, the ratio of these two numbers describes the extent to which the coders agree on the specific category. Mathematically this idea can be expressed and extended by using conditional probabilities.

The probability of an event A under the condition C is generally described by $P(A|C) = \frac{P(A \cap C)}{C}$. Applied to the current concept, we define event A_i as the case that all coders agree on category i. This means that all coders assign a coding unit to the same category. We define condition C_i as the case where at least one coder assigns a coding unit to category i. In **Figure 2**, event A_i

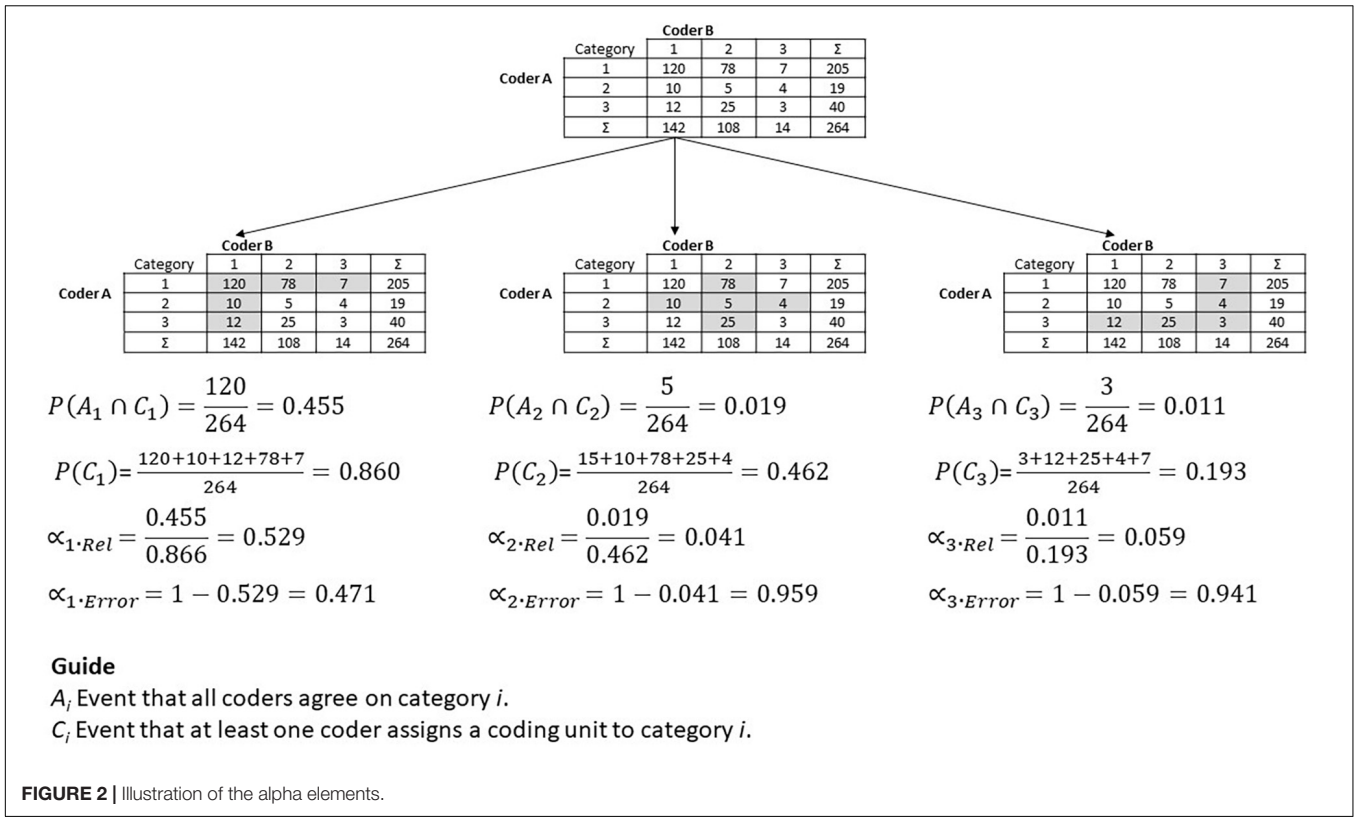


FIGURE 2 | Illustration of the alpha elements.

is the corresponding cell on the diagonal, with event C_i reflected by the gray cells for each category. With these definitions in mind, we can define the alpha reliability and the alpha error for category i as

$$\propto_{i,Rel} = \frac{P(A_i \cap C_i)}{P(C_i)} \quad (1)$$

$$\propto_{i,Error} = 1 - \frac{P(A_i \cap C_i)}{C_i} \quad (2)$$

The alpha error is the complementary probability of the alpha reliability. Equations 1 and 2 provide the central interpretation of the alpha elements. The alpha reliability is the probability that all coders agree on the category of a coding unit if at least one coder assigns the coding unit to that category. The alpha error is the probability that not all coders agree on the category of a coding unit if at least one coder assigns the coding unit to that category.

We suggest treating alpha reliability as an approximation of the probability that a coding unit of category i is classified as category i, and the alpha error as the probability that a coding unit of category i is not classified as category i. The reason for this interpretation of the conditional probabilities of the alpha elements is that the true category cannot be known. This interpretation of the alpha elements assumes that the assignment of a coding unit to this category by at least one coder is an adequate approximation for the amount of coding units "truly" belonging to that category. Furthermore, this interpretation of the alpha elements makes them comparable to the alpha errors used in significance testing.

Figure 2 shows the computations for an example where the alpha reliability for category one is 0.529. This means that the probability that a coding unit of category one is correctly classified as category one is about 53%. The same probability is about 4% for category two, and about 6% for category three. Here, a coding unit belonging to category two or three is only rarely classified as category two or three respectively. The alpha error for both of these categories is very high, with a probability of about 94–96%.

The occurrence of an alpha error means that a coding unit is wrongly assigned to another category. In this case, the data of the other categories will be biased as a result of errors in other categories. The beta elements account for these errors.

Beta Elements: Beta Reliability and Beta Error

A category's data is not only influenced by the alpha error of that category, but by errors in other categories as well. For example, a coding unit could be assigned to category one although it belongs to category two. When this occurs, the data of category one will be biased by errors made in category two. However, this error can only occur if an alpha error occurs in category two, meaning a coding unit truly belonging to category two is wrongly assigned to category one. The same influence can be expected for every other category.

This relationship can be mathematically expressed with conditional probabilities. The event E_j represents all cases where an alpha error of category j occurs. In **Figure 3**, this is illustrated

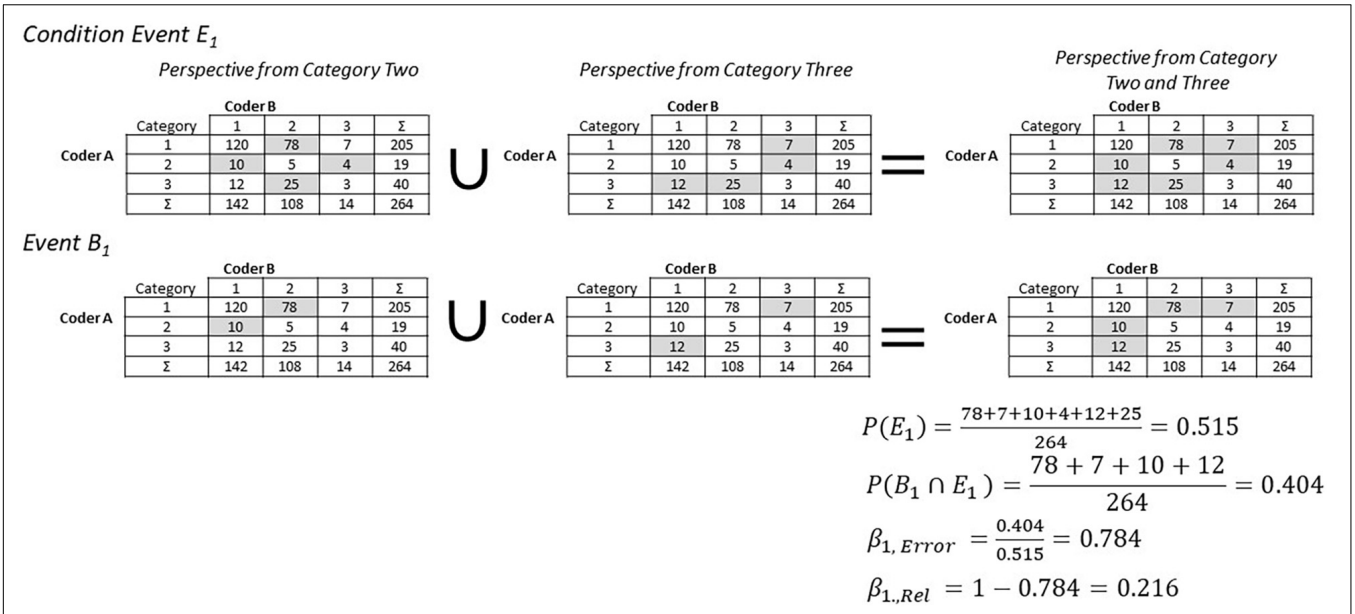


FIGURE 3 | Illustration of the beta elements.

by the gray cells for category two and three. Alpha errors of all other categories are relevant for estimating the beta error of category one. This situation is illustrated on the right side of **Figure 3**. The condition here for the beta error of category i is an occurrence of an alpha error in all other categories. In general, event E_i is defined as all cases where an alpha error occurs in all other categories except i.

$$E_i = \cup E_j, \ \ where\ i \neq j$$

To be relevant for category one, only those parts of the alpha errors of the other categories are relevant that guide coders to assign a coding unit to category one. This situation is illustrated in the second row of **Figure 3**. The corresponding event B_i represents all cases where at least one coder assigns a coding unit to category i, without the cases where all coders assign a coding unit to category i. The reason for the exclusion of the cases where all coders assign a coding unit to category i is that these cases do not represent an error. The beta error of category i is therefore defined as:

$$\beta_{i,Error} = \frac{P(B_i \cap E_i)}{P(E_i)} \tag{3}$$

Mathematical equation 3 can be simplified for computations by applying the concept of contemporary probabilities. As shown in the first row on the right side of **Figure 3**, $P(E_i)$ can be expressed as the complementary probability of the event that all coders agree on different categories (the diagonal of the table). Furthermore, as shown in the second row on the right side of **Figure 3**, $P(B_i \cap E_i)$ can be expressed by the complementary probability of the event that no coder assigns a coding unit to category i and that all coders assign a coding unit to category i (white cells).

Similar to the alpha elements, the beta reliability is the complementary probability to the beta error, describing the probability that no beta error will occur.

$$\beta_{i,Rel} = 1 - \frac{P(B_i \cap E_i)}{P(E_i)} \tag{4}$$

Using the example of **Figure 3**, the beta error for category one is 0.784. This means that the probability of assigning a coding unit to category one if an alpha error occurs in categories two or three is about 78%. The beta elements and the alpha elements offer the possibility to analyze the influence of errors in greater detail with the help of the assignment-error matrix (AEM).

The Assignment-Error Matrix

The assignment-error matrix is a tool for analyzing the influence of errors in one category on other categories. The diagonal cells show the alpha error for the specific category. The remaining cells describe the probability that an alpha error guides coders toward assigning a coding unit to another specific category. The interpretation of this matrix can best be explained using the example shown in **Table 1**. The alpha error for category one is

TABLE 1 | An example of an assignment-error matrix.

		Assigned Category		
True category	**Category**	**1**	**2**	**3**
	1	0.471	0.709	0.291
	2	0.690	0.959	0.310
	3	0.478	0.522	0.941

$\propto_{2,Error} = 0.959$
$\beta_{1,Error} = 0.787$
$\beta_{2,Error} = 0.860$
$\beta_{3,Error} = 0.353$
$AEM(2,1) = \frac{0.959*0.784}{0.959*(0.787+0.353)} \cong 0.690$

about 47%, i.e., in about 47% of the cases, a coding unit that truly belongs to category one is assigned to another category. When this error occurs, about 71% of the cases are assigned to category two, and about 29% of the cases to category three. Here, category two is more strongly impacted by the coding errors of category one than category three.

The alpha error of category two is about 96%, meaning that in about 96% of the cases, a coding unit truly belonging to category two is assigned to another category. When this error occurs, about 69% of the cases are assigned to category one, and 31% of the cases to category three. Here, category one is more strongly impacted by the coding errors in category two than category three.

The assignment-error matrix provides detailed information about how errors influence the data. With this example, category one and two are not well differentiated, meaning the development of the coding scheme should concentrate on creating better definitions and coding rules for distinguishing category one and two. In contrast, errors in category one and two do not strongly influence category three. If an alpha error occurs in category three, both remaining categories are impacted by this error in a similar way.

The values for the cells outside the diagonal can be easily estimated with the alpha and beta elements. The *condition* is that an alpha error occurs in the category under investigation, and that a beta error occurs in all other categories. The *target event* is that an alpha error occurs in the category under investigation, and a beta error in the other respective category. Equation 5 expresses this relationship.

$$AEM\left(i,j\right) = \frac{\propto_{i,Error} * \beta_{j,Error}}{\propto_{i,Error} * \sum_{j \neq i} \beta_{j,Error}} \tag{5}$$

The iota elements comprise the final aspect of this concept.

Iota Elements

The last part of this concept summarizes the different types of errors while correcting the values for chance agreement, providing the final reliability measure for every category. In a first step, the alpha error and the beta error have to be calculated under the condition of guessing. The concept here assumes that every coder randomly chooses a category, and that every category has the same probability of being chosen. The probability for every combination with k categories and c coders is $p = \frac{1}{k^c}$. The equations (1), (2), (3), and (4) introduced in Section "Alpha Elements: Alpha Reliability and Alpha Error" and "Beta Elements: Beta Reliability and Beta Error" can now be applied for the calculation of the corresponding values.

$$A_{i,Rel} = \frac{p}{1-\left(k-1\right)^c p} \tag{6}$$

$$A_{i,Error} = 1 - \frac{p}{1-\left(k-1\right)^c p} \tag{7}$$

$$B_{i,Error} = \frac{1-p*\left(k-1\right)^c - p}{1-k*p} \tag{8}$$

$$B_{i,Rel} = 1 - \frac{1-p*\left(k-1\right)^c - p}{1-k*p} \tag{9}$$

The chance corrected and normalized alpha reliability is

$$\alpha_i = \left| \frac{\propto_{i,Rel} - A_{i,Rel}}{1 - A_{i,Rel}} \right| \tag{10}$$

Please note that normalization means here that the values can only range between 0 and 1. Although the definition of α_i appears clear, the equation for β_i still has to be explained. The beta errors are designed in such a way that they describe how errors influence the data of the category under investigation *if* errors occur in the other categories. However, they do not provide direct information about the probability of a beta error occurring, meaning that the probability for the condition of the beta errors has to be estimated in a first step. As described in Section "Beta Elements: Beta Reliability and Beta Error," $P(E_i)$ represents the probability for the condition of beta errors, and can be expressed as the complementary probability of the event that all coders agree on the different categories (the diagonal of the table in **Figure 3**). For the beta error under the condition of guessing, the corresponding probability is $1 - k^*p$. The realized beta errors with chance correction are shown in Equation 11.

$$b_{i,Error} = P\left(E_i\right) * \beta_{i,Error} - (1-kp) * B_{i,Error} \tag{11}$$

The complementary probability represents the corresponding realized beta reliability as shown in Equation 12. Equation 13 represents the normalized beta reliability.

$$\begin{aligned} b_{i,Rel} &= (1 - P\left(E_i\right) * \beta_{i,Error}) - \left(1 - \left(1-kp\right) * B_{i,Error}\right) \\ &= \left(1-kp\right) * B_{i,Error} - P\left(E_i\right) * \beta_{i,Error} \end{aligned} \tag{12}$$

$$\begin{aligned} \beta_i &= \left| \frac{(1 - P\left(E_i\right) * \beta_{i,Error}) - (1 - \left(1-kp\right) * B_{i,Error})}{1 - (1 - \left(1-kp\right) * B_{i,Error})} \right| \\ &= \left| \frac{\left(1-kp\right) * B_{i,Error} - P\left(E_i\right) * \beta_{i,Error}}{(1-kp) * B_{i,Error}} \right| \\ &= \left| 1 - \frac{P\left(E_i\right) * \beta_{i,Error}}{(1-kp) * B_{i,Error}} \right| \end{aligned} \tag{13}$$

The utilization of the absolute value for α_i und β_i is inspired by the chi-square statistic in contingency analysis. The idea behind this approach is that the more a system is behind the observed data, the more data values deviate from a data set generated by random guessing. With this in mind, the final iota is defined as shown in Equation 14.

$$\mathrm{I}_i = \frac{\alpha_i + \beta_i}{2} \tag{14}$$

I_i can be roughly interpreted as the average probability that no error occurs. It is 1 in the case of no error, and 0 if the errors equal the amount of errors expected by guessing.

Iota describes the reliability of every single category. In some situations additional information on the reliability of the complete scale is necessary. In order to aggregate the single

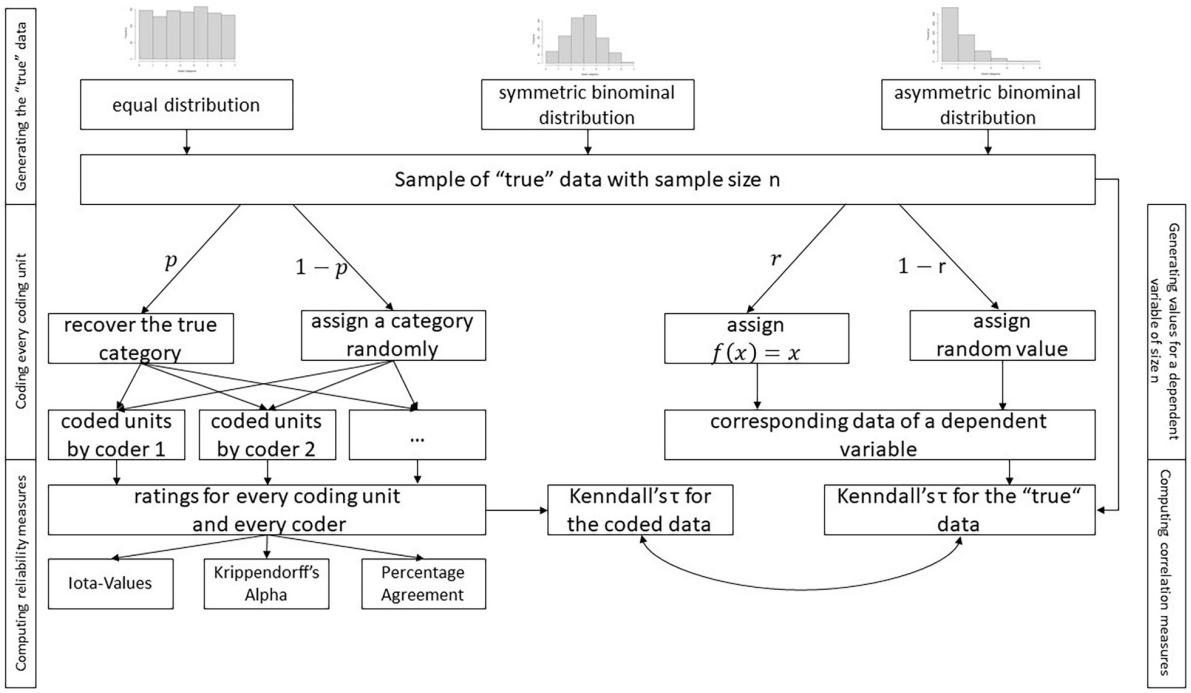


FIGURE 4 | Design of the simulation study.

iota values, the Iota Concept suggests the average iota and the minimal iota as possible indicators. The average iota represents the mean of all iota values taking all available information into account. This, however, implies the opportunity that the reliability is overestimated as a low reliability in one category can be compensated by a high reliability in other categories. This problem is addressed with the minimum iota using only the information of the category with the lowest reliability.

The following chapter presents the results of a simulation study aiming to generate cut-off values for the new reliability measure, and provides insight into its statistical properties.

SIMULATION STUDY OF THE NEW RELIABILITY CONCEPT

Simulation Design

A simulation study was conducted with *R* to provide an answer to the following questions:

(1) How strongly are the reliability values of the new concept correlated with the true reliability of a coding scheme?
(2) How does the distribution of the data influence the reliability values?
(3) How does the number of categories influence the reliability values?
(4) How does the number of coders influence the reliability values?
(5) How does the new measure perform in comparison to other reliability measures?
(6) Which cut-off values should be used for judging the reliability of a coding scheme?

A simulation study was performed to answer these questions. **Figure 4** shows the design of the simulation.

The first step generated coding units. For modeling the distribution of the categories of the coding units in the population, an equal distribution (probability for every category $1/k$), a symmetric binominal distribution (probability 0.5, size $k-1$), and an asymmetric binominal distribution (probability 0.2, size $k-1$) were used. For every distribution, a sample was drawn with different sample sizes n = 10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500. This procedure was repeated 50 times.

The coding process was simulated after generating samples of true data, i.e., every coding unit was coded by a coder who applied a coding scheme. The coding scheme guided a coder to recover the true category with the probability p. If the coder failed, a category was randomly assigned to the coding unit. To simplify the simulation, it was assumed that p was equal for each category. In the case of $p = 0$, there was no reliability, with a coder randomly assigning a category to a coding unit. The coding fluctuated unsystematically. In the case of $p = 0.99$, the coding scheme led a coder to assign the same category if the coding unit offered the corresponding indication. The coding systematically provided stable results. The value of p represented the reproducibility of the coding scheme and could be interpreted as true reliability. This process was repeated for different p values ranging from "0" to "0.99" and for every coder.

The coding of every coder provided the basis for computing different reliability measures. The new iota values, Krippendorff's

alpha, and the percentage agreement were applied in the current simulation. Krippendorff's alpha and percentage agreement provided comparison standards for the new measure. Percentage agreement represented a more liberal measure, and Krippendorff's alpha a more conservative one (Zhao et al., 2013). The average iota and the minimum iota were computed to generate a measure for the complete coding scheme. The process described above was repeated for up to eight categories and up to eight coders. This simulation helped answer questions 1–5.

A dependent variable was simulated in a similar way to answer question 6. The idea behind this attempt was that the cut-off value for judging the reliability of a coding scheme should consider the effects of further statistical computations and derived decisions. As a result, the correlation of the true data in a sample was compared to the correlation estimated based on the coded data. This attempt allowed the estimation of the expected deviation between the true and the observed correlation for different reliability values. The correlation was measured with Kendall's tau, which is applicable for ordinal data. As a result, this simulation focused only on ordinal data, using a simple relationship. The strength of the correlation was simulated with the probability *r*. The corresponding values for tau are outlined in **Supplementary Appendix B**, and the results are reported in the following sections.

TABLE 2 | Effect sizes of the impact of different factors in the reliability measures.

Factor	Average Iota	Minimum Iota	Krippendorff's Alpha	Percentage Agreement
	η	η	η	η
Observed Concentration	0.00	0.00	0.00	0.00
True Reliability (p)	0.87	0.84	0.90	0.74
Number of Categories (k)	0.00	0.01	0.00	0.03
Number of Coders (c)	0.05	0.03	0.00	0.15
Sample Size	0.00	0.00	0.00	0.00
Distribution	0.00	0.00	0.00	0.00
True Reliability: Categories	0.00	0.01	0.00	0.00
True Reliability: Coders	0.01	0.01	0.00	0.03
True Reliability: Sample Size	0.01	0.01	0.00	0.00
True Reliability: Distribution	0.00	0.00	0.00	0.00
Categories: Coders	0.00	0.00	0.00	0.00
Categories: Sample Size	0.00	0.00	0.00	0.00
Categories: Distribution	0.00	0.00	0.00	0.00
Coders: Sample Size	0.00	0.00	0.00	0.00
Coders: Distribution	0.00	0.00	0.00	0.00
Sample Size: Distribution	0.00	0.00	0.00	0.00

Results of the Simulation Study

Results on the Scale Level

A data set of 808,500 cases was generated. **Table 2** shows the results of an ANOVA focusing on the effect sizes. According to Cohen (1988), an η^2 of at least 0.01 represents a small effect; of at least 0.06 a medium effect; and of at least 0.14 a strong effect.

About 87–90% of the variance can be explained by the true reliability for the average iota and Krippendorff's Alpha. The true reliability can explain about 84% of the variance of the minimal iota values. Average iota, minimum iota, and Krippendorff's alpha show a very strong relationship with the true reliability, and are able to provide an adequate indication of it. In contrast, the true reliability can only account for about 74% of the variance of the percentage agreement; percentage agreement is more problematic than the other measures since it may be influenced by construct irrelevant sources.

Whereas Krippendorff's alpha is not influenced by any other source of variance (e.g., the number of categories or the number of coders), the number of coders influences average iota. However, this effect is very small, with an η^2 of 0.05, making it practically not important. Minimum iota shows a small bias with respect to the number of categories, with an η^2 of 0.03, which is also of minimal practical relevance. In contrast, the number of coders heavily influences the reliability estimation by the percentage agreement, with an η^2 of 0.15. Thus, the values for percentage agreement are not comparable across coding with a different number of coders.

The simulated distributions, the sample size, and the number of categories do not bias the values of Krippendorff's alpha, the average iota, and the minimum iota. In contrast, percentage agreement is influenced by the number of categories, but not by the sample size. However, this effect is very small.

Figure 5 shows the estimated marginal means for the different configurations of the true reliability and the deviation of the estimated values from the true reliability. It becomes clear that no measure stands in a linear relationship with the true reliability; all measures underestimate this. Average iota, minimum iota, and percentage agreement show the highest degree of underestimation near 0.75, while Krippendorff's alpha shows the maximum deviation near 0.50. In this sense, all measures can be classified as rather conservative.

Polynomial functions from degree one to four are calculated to describe the relationship between the reliability measures and the expected deviation of Kendall's tau. **Table 3** reports the R^2 values for the different functions to select an appropriate model.

R^2 increases when *r* increases, regardless of which performance measure is under investigation. This means the impact of reliability on the data is more important in situations where a strong relationship exists than in situations where there is only a weak relationship. For the average and minimum iota, a polynomial function of degree two accounts for more variance as a linear function (degree one). However, polynomials of degree three and four do not noticeably improve the R^2. The relationship between the iota measures and the deviation can therefore be characterized best with a polynomial of degree two. In contrast, Krippendorff's alpha and percentage agreement can be best characterized by a linear relationship.

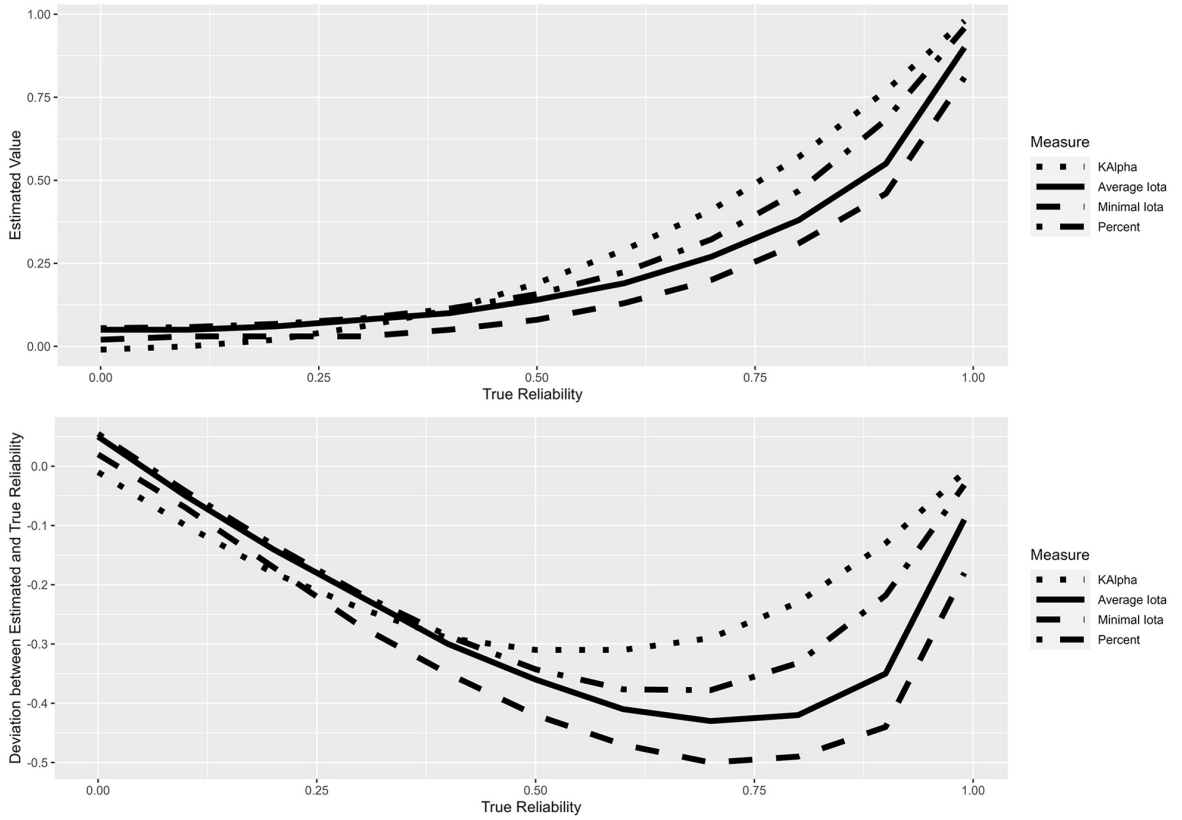


FIGURE 5 | Estimated marginal means of the measures and their deviation from true reliability.

Figure 6 shows the polynomials for the different reliability measures describing the expected deviations from the true correlation within a sample. The horizontal lines in **Figure 6** show where the deviation between the estimated expected Kendall's tau, and the true Kendall's tau is 0.20. This information can be used to derivate cut-off values for judging the quality of a coding scheme. If a researcher allows an expected deviation of at most 0.20 between the true and the estimated Kendall's tau, the average iota should be at least 0.474, the minimum iota should be at least 0.377, Krippendorff's alpha at least 0.697, and percentage agreement at least 0.711. This can be seen by the intersection of the horizontal line for 0.20 and the curve for $r = 1.0$.

Results on the Categorical Level

An ANOVA was performed to describe the relationship between the true reliability and the estimated iota values on the level of single categories. The effect sizes eta and omega are: true reliability (p): 0.84, number of categories (k): 0.00, number of coders (c): 0.03, true reliability: categories: 0.00, true reliability: coders: 0.01, and categories: coders: 0.00. First, the true reliability is the central source of variance for iota on a categorical level. It explains about 84% of the variance. Iota is thus a strong indicator of the reliability on the categorical level. Only the number of coders slightly influences iota, but according to Cohen (1988), only with a minor effect.

In order to describe the relationship between the true reliability and the caused iota values, several functions are fitted to the data. The function $f(x) = x^{3.861705}$ reveals a residual standard error of 0.1231 by 3,891,774 degrees of freedom. This function has the advantage that it comprises the extreme points of the scale "zero" and "one," which is why this function is used for further modeling: it is invertible in the necessary range of values. The inverse function is:

$$f(x)^{-1} = \sqrt[3.861705]{x}$$

Applying this inverse function on iota will produce linearized iota values which allow an interpretation as probabilities. Based on the new measure, the following chapter analyzes the performance and configuration of AI in the context of business education.

ANALYZING THE PERFORMANCE AND CONFIGURATION OF ARTIFICIAL INTELLIGENCE

Simulation Design

Several algorithms of AI exist to analyze textual data. The current study focuses on decision tree-based algorithms and neural nets; these two kinds of AI show different characteristics.

Decision trees are well-suited for classification tasks and have the advantage that the results are understandable for people (Lanquillon, 2019; Richter, 2019). This is a very important feature because the results of a learning analytics application should be understood by students and educators as they foster confidence in the recommendations derived. Understanding the way an AI produces a result is also crucial within a legal context whenever the results provided by the software are used for decisions that potentially have a strong impact on the further education of students. Although neural nets are very powerful concepts of AI, understanding the transformation from input data to output data is more difficult. In the current study, the concept of decision trees is implemented using the packages *rpart* (Therneau et al., 2019) and *ranger* (Wright and Ziegler, 2017). To realize neural nets, the study uses the package *nnet* (Venables and Ripley, 2007). The current study analyzes the performance of these three implementations in an attempt to find hyperparameter configurations optimizing their performance. **Figure 7** presents the corresponding research design.

TABLE 3 | Modeling the relationship of different reliability measures and the absolute deviation for Kendall's tau.

Measure	*r*	R^2			
		Polynomial Degree 1	Polynomial Degree 2	Polynomial Degree 3	Polynomial Degree 4
Average Iota	0.00	0.068	0.069	0.072	0.076
	0.10	0.093	0.093	0.095	0.097
	0.20	0.153	0.157	0.157	0.157
	0.30	0.227	0.241	0.241	0.241
	0.40	0.303	0.329	0.332	0.332
	0.50	0.374	0.412	0.415	0.415
	0.60	0.440	0.488	0.492	0.493
	0.70	0.501	0.557	0.561	0.561
	0.80	0.556	0.617	0.621	0.621
	0.90	0.604	0.669	0.673	0.673
	1.00	0.646	0.714	0.716	0.716
Minimum Iota	0.00	0.080	0.082	0.083	0.086
	0.10	0.103	0.107	0.108	0.110
	0.20	0.156	0.168	0.168	0.168
	0.30	0.221	0.244	0.244	0.244
	0.40	0.288	0.324	0.326	0.326
	0.50	0.352	0.401	0.405	0.405
	0.60	0.414	0.475	0.481	0.481
	0.70	0.472	0.545	0.552	0.553
	0.80	0.525	0.609	0.617	0.618
	0.90	0.573	0.666	0.675	0.677
	1.00	0.617	0.717	0.726	0.728
Krippendorff's Alpha	0.00	0.101	0.103	0.111	0.111
	0.10	0.131	0.132	0.138	0.138
	0.20	0.198	0.200	0.203	0.203
	0.30	0.281	0.285	0.285	0.286
	0.40	0.368	0.373	0.373	0.375
	0.50	0.452	0.457	0.458	0.460
	0.60	0.533	0.540	0.542	0.545
	0.70	0.609	0.618	0.620	0.624
	0.80	0.680	0.689	0.693	0.697
	0.90	0.744	0.755	0.759	0.764
	1.00	0.802	0.814	0.818	0.824
Percentage Agreement	0.00	0.073	0.074	0.077	0.077
	0.10	0.095	0.095	0.097	0.098
	0.20	0.145	0.145	0.148	0.149
	0.30	0.209	0.209	0.211	0.212
	0.40	0.276	0.277	0.279	0.281
	0.50	0.342	0.344	0.346	0.348
	0.60	0.407	0.409	0.412	0.414
	0.70	0.469	0.472	0.475	0.477
	0.80	0.526	0.531	0.534	0.537
	0.90	0.578	0.584	0.587	0.590
	1.00	0.626	0.632	0.636	0.639

The simulation study is based on real empirical textual data which was analyzed in several studies. **Table 4** provides an overview of the different data sets. A detailed list of the inter-coder reliability can be found in **Supplementary Appendix A**. Every data set is divided into training and evaluation data. 75% of the complete data is used for training, and the remaining data for evaluation. AI performance can be tested here with textual data that is unknown by AI. The iota concept, Krippendorff's Alpha, and percentage agreement are used for performance evaluation. Data splitting is repeated 30 times by applying stratified custom sampling.

A numerical representation of the texts was created based on the training data of a sample. Here, the texts were transformed into a document-term matrix (DTM) showing the documents in the rows and the frequency of the words in the columns (bag-of-words approach). This was done by applying the package *quanteda* (Benoit et al., 2018). The words were reduced to nouns, verbs, adjectives, and adverbs, helping reduce the dimension of the DTM, and limiting the analysis to the words carrying the most semantic meaning (Papilloud and Hinneburg, 2018). The words were also lemmatized. These steps were performed with *UDPipe* (Straka and Straková, 2017; Wijffels et al., 2019), using the *HDT-UD 2.5* created by Borges Völker et al. (2019).

In a next step, the words were filtered with the two approaches of joint mutual information maximization (JMIM) (Bennasar et al., 2015) and information gain, each provided by the *praznik* package (Kursa, 2021). With the help of these filters, the number of words was reduced to 5, 10, 15, 20, and 25% of the initial number. This step was very important for neural nets in light of how they typically have the curse of dimensionality.

The training of the different forms of AI was conducted based on the filtered DTM. The data was here again divided into training data and test data to perform hyperparameter tuning, with the aim to find the best configuration for the different algorithms. The hyperparameter tuning used 50 custom samples of training and test data. 75% of the data was for training, and the remaining part for testing. The hyperparameter tuning was done with random search (Bergstra and Bengio, 2012) because it was not clear which hyperparameters were the most important for analyzing didactical and pedagogical texts. **Table 5** reports the standard configuration and the search space for the different parameters. A description of the meaning of the

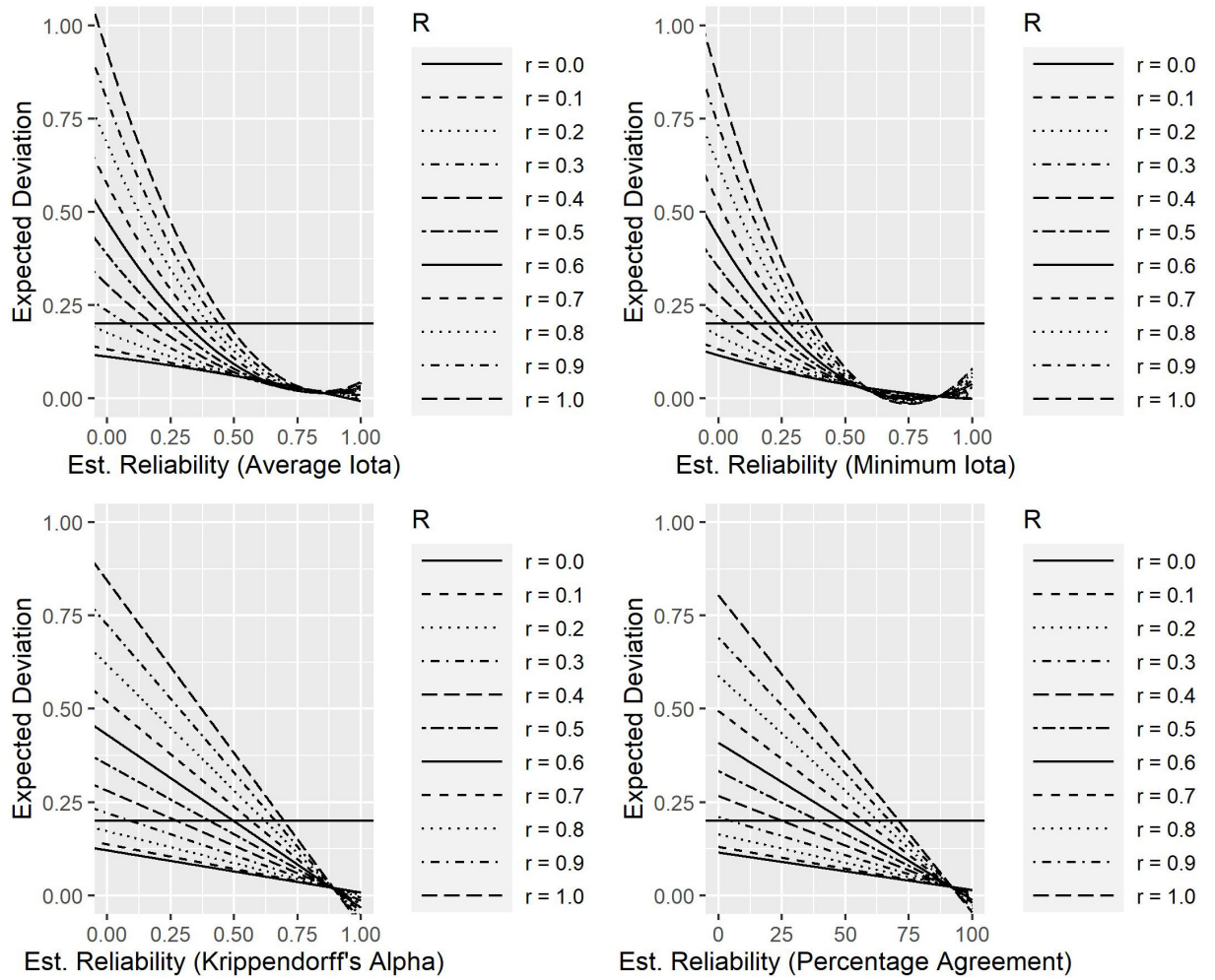


FIGURE 6 | Reliability values and expected deviation from true correlation.

different parameters can be found in the documentation of the applied *R* packages.

A central problem for most algorithms of AI is that they achieve good performance for categories with a high frequency, and low performance for categories with a low frequency (Haixiang et al., 2017). This is problematic in the context of learning analytics, because extreme characteristics of relevant learning concepts imply individualized learning processes, even though these extreme characteristics usually have a low frequency. For example, underachievers and overachievers need individual learning processes to fully develop their potential. However, this requires a reliable diagnosis of characteristics. Different approaches exist to solve this problem of imbalanced data. The current study applied an oversampling strategy where artificial data sets were generated to balance the frequencies of the different categories. According to Haixiang et al. (2017), this approach should be used if the frequencies of some categories are very small and can be implemented using the synthetic minority oversampling technique (SMOTE). The relevant parameters for SOMTE were also added to the hyperparameter tuning. All computations were done with the *mlr3* interface (Lang et al., 2019). The following section reports the results.

Results

An ANOVA was performed using the SPSS software to generate first insights. **Table 6** reports the effect sizes for the different factors. A detailed list of the achieved performance measures for every construct can be found in **Supplementary Appendix A**.

About 90% of the variation in the percentage agreement and the average iota is explained by the factors shown in **Table 6**. In contrast, the investigated configuration explains about 87% of the variation of minimum iota, and only 78% of Krippendorff's Alpha. In each case, it depends on the operationalization of the construct under investigation, as this is the most important factor for explaining the performance of AI. The construct explains at least 72% of the total variation. Thus, the configuration of AI only slightly affects its performance. The AI configuration explains between 3.6% of the total variation of the percentage agreement, and up to 7.8% of the minimum iota.

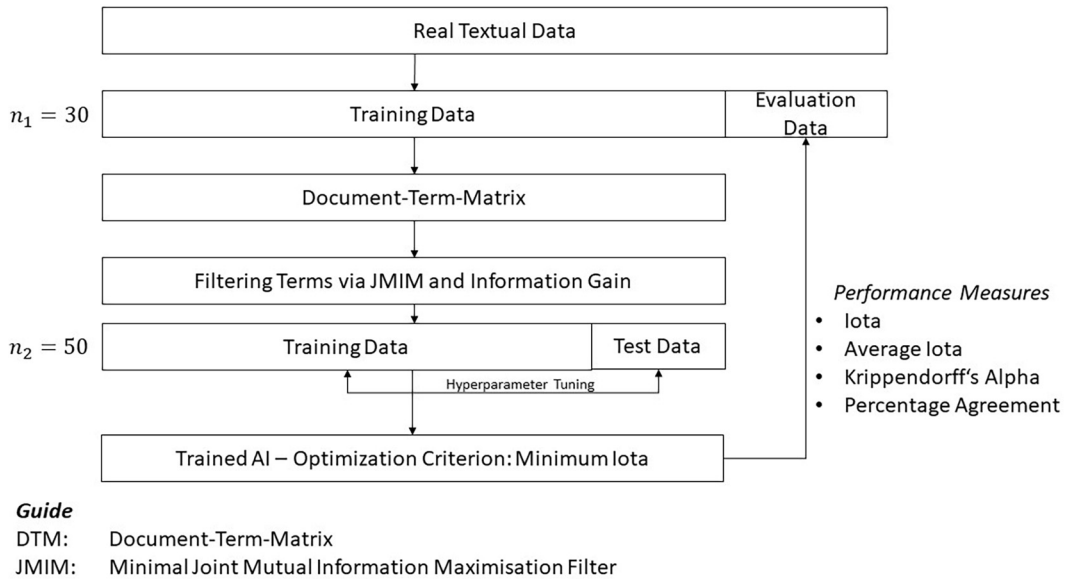


FIGURE 7 | Research design for testing AI performance.

TABLE 4 | Empirical data for the simulation.

#	Concept/Model/Label	# Constructs	# Categories	Sample size	Kind of text	Characteristics of writers	Source
Texts produced by apprentices, students of business administration, and pre-service teachers for business education							
1a	Basic ideas of expenses	9	2	632	Written essays	Apprentices and students (EQR-Level 4–7)	Berding, 2019; Berding and Jahncke, 2020
1b	Formal strategies for expenses	5	2	632	Written essays	Apprentices and students (EQR-Level 4–7)	Berding, 2019; Berding and Jahncke, 2020
2a	Basic ideas of earnings	8	2	640	Written essays	Apprentices and students (EQR-Level 4–7)	Berding, 2019; Berding and Jahncke, 2020
2b	Formal strategies for earnings	5	2	640	Written essays	Apprentices and students (EQR-Level 4–7)	Berding, 2019; Berding and Jahncke, 2020
3	Basic ideas of capital, equity capital, and debt capital	16	2	149	Written essays	Students (EQR-Level 6–7)	Berding et al., 2021
4	Basic ideas of costs and performance	11	2	112	Written essays	Students (EQR-Level 6–7)	Berding et al., 2021
5	Self-reflection competence	3	4	265	Written essays	Students (EQR-Level 6)	Jahncke, 2019
6	Quality of lesson plans	3	4-5	455	Written lesson plans	Students (EQR-Level 7)	Riebenbauer, 2021
Texts representing learning materials in business education							
7	Quality of learning tasks in accounting education	14	2–3	1,707	Textbook tasks for apprentices		Berding et al., 2021; Kühne, 2021
8	Quality of learning tasks for sustainable business administration	7	2–3	1,468	Textbooks tasks for apprentices		Slopinski et al., in preparation
9	Sustainable Development Goals (SDGs)	9	2	435	Instructional textbook texts and tasks for apprentices		Slopinski et al., in preparation

Surprisingly, the operationalization of a construct is more important for the percentage agreement and the average iota than for the minimum iota and Krippendorff's Alpha. Krippendorff's Alpha is the least influenced by the constructs under investigation. In this context, operationalization means the quality of how a construct is defined and described in the coding scheme of a content analysis.

Shifting the focus from the total variation to the variation within a construct ("ETA Square Within"), there is a clear impact of the algorithm on determining AI. The main effect of the algorithm varies from 1% for Krippendorff's Alpha to 21% for average iota. In some cases, the interaction between the construct and the algorithm is more important than the main effect. For example, the interaction explains about 24% of the within variation for minimum iota, while the main effect explains only 16%. The other configurations are less important. Again, Krippendorff's Alpha is least influenced by the different options for the configuration of AI.

A three-level structural equation model was computed with MPlus 8.6 using the Bayes estimation to generate more detailed

TABLE 5 | Hyperparameter configuration and search space.

rpart				Ranger				Nnet			
param	***S***	**Search space**		**param**	***S***	**Search space**		**param**	***S***	**Search space**	
		Min	**Max**			**Min**	**Max**			**Min**	**Max**
cp	0.01	0	0.01	replace	True	True	False	decay	0	0	0.2
maxdepth	30	25	30	maxdepth	30	25	90	size	5	2	20
minbucket	7	1	5	splitrule	gini	gini, extratrees					
minsplit	20	1	5								
dup_size	1	1	5	dup_size	1	1	5	dup_size	1	1	5
smote.k	1	1	6	smote.k	1	1	6	smote.k	1	1	6

S, standard, param, parameter.

TABLE 6 | Effect sizes of the influence of different factors and the achieved performance measures.

Factor	ETA Square				ETA Square Within			
	Minimum Iota	**Average Iota**	**Kalpha**	**Percent**	**Minimum Iota**	**Average Iota**	**Kalpha**	**Percent**
Algorithm	0.027	0.027	0.002	0.016	0.159	0.213	0.008	0.194
Algorithm * Construct	0.041	0.028	0.033	0.013	0.239	0.219	0.140	0.160
Filter	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.006
Filter * Construct	0.003	0.002	0.009	0.002	0.016	0.014	0.037	0.021
Filter Percentage	0.000	0.000	0.001	0.000	0.001	0.001	0.002	0.000
Filter Percentage * Construct	0.002	0.001	0.003	0.001	0.011	0.009	0.013	0.007
Tuned	0.000	0.000	0.006	0.000	0.000	0.000	0.027	0.000
Tuned * Construct	0.000	0.000	0.005	0.000	0.000	0.000	0.022	0.000
Algorithm * Filter	0.000	0.001	0.000	0.001	0.002	0.004	0.002	0.007
Algorithm * Filter * Construct	0.003	0.003	0.004	0.002	0.020	0.020	0.017	0.019
Algorithm * Filter Percentage	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000
Algorithm * Filter Percentage * Construct	0.002	0.002	0.002	0.001	0.014	0.014	0.010	0.012
Filter * Filter Percentage	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.001
. . .								
Construct	0.787	0.844	0.719	0.901	–/–	–/–	–/–	–/–
Total Eta Square	0.869	0.909	0.801	0.938	0.474	0.507	0.350	0.439

Only factors with a relevant eta square are shown.

The column "Eta Square" represents the proportion of the total variation that a factor explains.

insights into the configuration of AI. In the current case, a multi-level modeling approach is more appropriate because the generated data is nested within construct and sample selections (see **Figure 7**). As Wang and Wang (2020) summarize, Bayes estimation has many advantages. The most important ones are that models can include both categorical and continuous data, that estimation of complex models is possible, and that this kind of estimation prevents problematic solutions (e.g., negative residual variances). **Table 8** reports these findings.

As the values for R^2 indicate, the hyperparameter tuning does not explain much of the variation of the different performance measures. In most cases, the application of the filter method "information gain" leads to decreased performance values, meaning that JMIM is the superior filter method. Regarding the number of features included in the training, most coefficients are negative. This means that including a smaller number of words leads to an increased performance for all three algorithms. The following section discusses the approach, results, and implications.

TABLE 7 | Example for assignment-error-matrices for different Sub-Groups.

		Assigned Category					
		All participants		**Men (*n* = 73)**		**Women (*n* = 71)**	
		0	**1**	**0**	**1**	**0**	**1**
True Category	0	0.134	1.00	0.143	1.00	0.127	1.00
	1	1.00	0.390	1.00	0.320	1.00	0.5

This is only an example for illustration based on one iteration of the underlying sample.
Six people did not provide information on gender.

DISCUSSION

Learning analytics is an emerging technology that supports stakeholders in the improvement of learning and teaching (Larusson and White, 2014; Rienties et al., 2020). The current

TABLE 8 | Standardized coefficients for decision trees (rpart), RandomForest (ranger), and neural net (nnet).

N	Optimization	Evaluation			
	1,350,000	27,000			
Measure	**Minimum Iota**	**Minimum Iota**	**Average Iota**	**Krippendorff's Alpha**	**Percentage Agreement**
Decision trees (rpart)					
R^2	0.004	0.009	0.008	0.008	0.001
Filter	0.002	0.085*	0.075*	–0.077*	0.015*
Filter Percentage	–0.017*	–0.037*	–0.035*	0.015*	–0.023*
cp	0.026*	0.005	0.003	0.026*	0.002
maxdepth	–0.001	–0.002	0.005	0.008	0.002
minbucket	0.035*	0.014*	0.014*	0.005	0.012
minsplit	0.003*	–0.002	–0.003	0.009	–0.006
Dup size	–0.040*	–0.009	–0.008	–0.022*	–0.007
Smote K	–0.003*	0.011	0.008	0.019*	0.003
Filter: 0 = jmim; 1 = information gain					
RandomForest (ranger)					
R^2	0.027	0.002	0.003	0.017	0.008
Filter	–0.152*	–0.015*	–0.035*	0.011	–0.077*
Filter Percentage	–0.005*	–0.037*	–0.034*	–0.084*	–0.016*
Replace	0.001	–0.007	–0.005	–0.001	–0.001
splitrule	–0.014*	0.010	0.013*	0.080*	0.020*
maxdepth	0.015*	0.002	0.004	0.037*	0.002
Dup size	–0.060*	–0.003	0.001	0.013	0.003
Smote K	0.008*	–0.002	0.006	0.037*	0.015
Filter: 0 = jmim; 1 = information gain Replace: 0 = false; 1 = true splitrule: 0 = gini; 1 = extra trees					
Neural net (nnet)					
R^2	0.075	0.007	0.016	0.013	0.034
Filter	–0.258*	–0.069*	–0.120*	–0.066*	–0.182*
Filter Percentage	0.013*	–0.044*	–0.036*	–0.081*	–0.014*
Decay	0.053*	–0.011	–0.015*	0.048*	–0.010
Size	0.003*	0.006	0.007	–0.009	0.002
Dup size	–0.074*	0.006	0.006	–0.001	0.002
Smote K	0.009*	0.002	0.000	0.008	0.001
Filter: 0 = jmim; 1 = information gain					

state of that technology uses data from different sources providing valuable knowledge and recommendations (Ifenthaler and Widanapathirana, 2014; Liu et al., 2018; ElSayed et al., 2019). However, the currently used kinds of data only represent students' learning actions on a surface-level and provide only a limited insight into students' cognition and motivation (Reich, 2015). Textual data can close this gap and further increase the value of learning analytics for learning and teaching by providing a deeper insight into students' knowledge, concepts, attitudes, and beliefs.

Realizing this potential requires the application of AI, since learning analytics applications have to understand and to interpret textual data in order to generate valuable knowledge based on scientific models and theories (Wong et al., 2019; Luan et al., 2020). In other words, AI has to conduct parts of a content analysis with a sufficient accuracy as the interpretation leads to corresponding interventions and recommendations. This paper has developed an original contribution to the field of content analysis and its application with AI in several forms:

(1) Previous measures often used in content analysis such as Krippendorff's Alpha, percentage agreement, Scott's Pi, and Cohen's Kappa (Lovejoy et al., 2016) are based on the basic ideas of classical test theory and describe the reliability of a scale with one single numeric value assuming that the reliability is constant for the complete scale (Feng and Zhao, 2016). The Iota Concept is based on the basic ideas of modern test theory (de Ayala, 2009; Baker and Kim, 2017; Bonifay, 2020; Paek and Cole, 2020) and provides a measure for every category and for the complete scale allowing a deeper insight into the quality of content analysis. Furthermore, the new Iota Concept provides a gate to apply other tools developed in item response theory for content analysis (see theoretical implications for more details).
(2) The previous measures are based on problematic assumptions as Zhao et al. (2013) worked out. The Iota Concept avoids these problematic assumptions since it is based completely on the mathematical

concept of conditional probabilities which allows a clear interpretation. Of course, the basic assumptions have to be discussed in further research. For example, the current version of iota assumes complete randomness as a kind of random selection with repetition. This could be problematic as complete randomness does not occur in practice (Zhao et al., 2013). However, the Iota Concept provides other measures that do not make a chance correction and thus avoid this problematic assumption. Thus, false conclusions can be avoided with the help of the new concept.

(3) Besides contributions to a progression in the field of content analysis, the current study offers insights in how well AI can interpret textual data from educational contexts and how the judgment of the quality depends on the chosen measure of reliability (see theoretical implications for more details). For practical applications this paper offers suggestions for the optimal configuration of AI that save researchers and users of AI both time and costs (see practical implications for more details).

(4) The Iota Concept can be used to evaluate possible bias in the recommendations of AI-supported learning technologies. Thus, this concept contributes to fill a gap identified by Luan et al. (2020). They determined that AI can reproduce bias and disadvantages minorities. With the help of the assignment-error-matrix these systematic errors can be discovered (see theoretical implication for an example).

In comparison to Krippendorff's Alpha, the new iota concept captures a similar amount of true reliability (84 and 87% in comparison to 90%) on a scale level. The main advantage of this new concept is that it provides reliability estimates for every single category. Here, iota is determined to be 84% of the true reliability. Similar to Krippendorff's Alpha, iota is not biased by the number of coders, the number of categories, the distribution of the data, or the sample size. As a consequence, it can be considered an adequate performance measure for inter-coder reliability.

Another advantage is that this new measure is based on less problematic assumptions (for details, see Zhao et al., 2013). Although the equations for α_i, $\alpha_{i,Rel}$, β_i, and $\beta_{i,Rel}$ appear similar to equations 3 and 6 in Zhao et al. (2013), the definition of its components is different. For example, $\alpha_{i,Rel}$ compares the number of units where all coders agree on with the number of all units of that category. This conceptualization prevents paradox 3 of "comparing apples with oranges" (Zhao et al., 2013). In the current study, only a few cases show results that can be clearly described as paradox, as **Supplementary Appendix A** shows. For example, the construct "validate" of the content analysis of tasks in accounting textbooks achieves a Krippendorff's Alpha near zero, and a percentage agreement of about 99%. The reliability estimates of every single category with iota show that both categories are measured reliably.

Surprisingly, Krippendorff's Alpha is the least influenced by the different constructs (72%), whereas percentage agreement is most influenced (90%) by them. Average iota and minimum iota land in between. Intuitively, a strong influence of the constructs should be seen as a good characteristic of a reliability measure, as it reflects how sensitive the measure is for the operationalization of the constructs. The same results occur for the within-subject factors. The different configurations can explain about 35% of the within-subject variation for Krippendorff's Alpha, and between 44 and 51% for the remaining measures. As the different configurations lead to different predictions of AI, a performance measure should be sensitive to the configuration. The new iota concept as a result can help to understand how different configurations of AI affect data.

The simulation study also provides first insights into meaningful cut-off values for different measures. By applying **Figure 6**, researchers can determine which amount of reliability is at least necessary for their study: **Figure 6** provides an estimation of the expected deviation between the true and the estimated sample correlation. If a researcher is interested in accurate results, the necessary reliability value can be defined. For example, the results of this simulation study show that the proposed cut-off value for Krippendorff's Alpha of at least 0.67 results in an expected deviation of 0.225, and the recommended cut-off value of 0.800 leads to an expected deviation of 0.105 (Krippendorff, 2019). Cohen (1988) does not explicitly develop effect sizes for Kendall's tau, although he does describe a classification system where the impact of correlations changes every 0.20 units (lower 0.10: no practical relevant effect, 0.10 to lower 0.30: small effect, 0.30 to lower 0.50: medium effect, 0.50 and above: strong effect). An Alpha of at least 0.67 ensures that the deviation has only a

TABLE 9 | Cut-off values for different measures, and number of constructs that reach the different cut-off values.

	Cut-Off Values	
Measure	**Maximum Deviation 0.20**	**Maximum Deviation 0.10**
Average Iota	0.474	0.601
Minimum Iota	0.377	0.478
Krippendorff's Alpha	0.697	0.805
Percentage Agreement	71.132	82.903
Number of constructs (rpart)		
Measure	Maximum Deviation 0.20	Maximum Deviation 0.10
Average Iota	70	55
Minimum Iota	67	58
Krippendorff's Alpha	12	2
Percentage Agreement	79	58
Number of constructs (ranger)		
Measure	Maximum Deviation 0.20	Maximum Deviation 0.10
Average Iota	77	60
Minimum Iota	73	63
Krippendorff's Alpha	13	2
Percentage Agreement	80	63
Number of constructs (nnet)		
Measure	Maximum Deviation 0.20	Maximum Deviation 0.10
Average Iota	64	46
Minimum Iota	65	44
Krippendorff's Alpha	9	2
Percentage Agreement	73	52

small practical effect; and of at least 0.800, no practical effect. Similar results can be derived accordingly for the other measures, as shown in **Table 9**.

The performance of AI can be discussed based on the cut-off values for the different reliability measures. Based on **Supplementary Appendix A**, **Table 9** reports the number of constructs that reach the different cut-off values.

According to **Table 9**, only 9–13 out of 90 constructs reach the minimal level for Krippendorff's Alpha. The recommended reliability level is only reached by two constructs. In contrast, between 73 and 79 out of 90 constructs achieve the cut-off values according to the percentage agreement. The evaluation of the AI performance for content analysis therefore largely depends on the chosen reliability measure. This finding is in line with the results generated by Hove et al. (2018) who found that different measures produce different numeric values for the same data.

As shown in this study, iota recovers an amount of reliability similar to Krippendorff's Alpha, is not practically influenced by other sources of variance, and relies on less problematic assumptions. The results of the new measure therefore appear more valid. According to average iota, between 64 and 70 constructs, and according to minimum iota, between 65 and 73 constructs achieve the minimal reliability requirements. In particular, minimum iota ensures that every single category is measured with a minimum degree of reliability. Based on this measure, AI can provide useful information about students' learning by analyzing textual data. The following section derives theoretical and practical implications of these findings.

CONCLUSION

Theoretical Implications

The Iota Concept provides a first step in the application of item response theory concepts to content analysis by providing a reliability measure for each category. Further research can build upon this approach and transfer further analytical tools to content analysis. From the different measures provided by the Iota Concept the assignment-error-matrix seems to be very promising. This matrix describes how coding units belonging to different true categories are assigned by coders to a specific category. Thus, this matrix represents how the data is generated.

Since the assignment-error-matrix characterizes the functionality of a coding scheme it can be used in the context of learning analytics to characterize if a content analysis produces similar data for different groups of people. In item response theory this problem is describes with the term "subgroup invariance" (e.g., Baker and Kim, 2017). Further research can address this idea for content analysis by developing corresponding significance tests.

As Seufert et al. (2021) found, at least two challenges occur when using AI for educational purposes. Firstly, AI may become so complex that humans are unable to understand the results generated. Secondly, AI may reproduce a bias which is part of a data set. As a result of these challenges, AI-literacy – defined as a "set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace [italic in the original]" (Long and Magerko, 2020) – includes the ability to understand how AI processes data and generates implications (Long and Magerko, 2020). The assignment-error-matrix can address both challenges since this matrix describes the data generation process. This idea can be illustrated with the following example based on the dataset from Berding and Jahncke (2020).

The dataset comprises 450 essays written by apprentices of business education. The corresponding coding scheme includes a scale for assessing whether the students acquired the concept that "expense" in accounting means that values are used for creating products and services, or not. After training an AI with the data from 300 participants, AI should assign the categories for the remaining 150 students. Based on these coding and the coding of a human coder, **Table 7** reports the resulting assignment-error-matrix.

As can be seen in **Table 7**, the alpha error is relatively low for both categories. Regarding the different sub-groups of men and women, the assignment-error-matrices differ. For example, the alpha error for the women in category 1 (concept is acquired) is about 0.18 percentage points higher than for the men. Thus, the coding scheme guides human coders more often to assign the texts of women to category 0 (concept not acquired) although the text truly indicates the acquisition of that concept. This bias in reproduced by the AI with the consequence that women are not correctly represented in the data. Furthermore, the data generation underestimates the performance of women in comparison to men. This can lead to false conclusions in research studies or biased recommendations in learning analytics applications.

Referring to the AI literacy of Long and Magerko (2020), the assignment-error-matrix could be a tool that is easy to interpret for understanding how AI may be biased and to foster the AI literacy of students. Furthermore, the assignment-error-matrix can help mitigate the problem of bias in learning analytic applications which currently remains a great challenge for that technology (Seufert et al., 2021).

The requirements for a reliable assessment of students' characteristics for learning analytics can be further discussed from another perspective. If the results generated by AI are used for judging the qualifications of learners, the demand for objectivity, reliability, and validity must be very high (Helmke, 2015), as errors can dramatically affect the educational path of learners. If the results are used only for fostering individual learning processes, the standards can be lower because the results provide orientation for teachers and educators in daily practice (Helmke, 2015). In daily practice, a high precision is not important as long as the direction of the conclusion leads to the right decisions (Weinert and Schrader, 1986). Here, the sign is more important than the concrete value. Thus, for fostering learning processes, less strict cut-off values are sufficient. Further studies should address which level of reliability is necessary for learning analytics applications to support individual learning (to be sure, the reliability of scientific studies has to adhere to higher standards).

Practical Implications

By providing information on every single category, developers of coding schemes gain orientation whenever a coding scheme needs revision. This allows a straighter process of development, can reduce costs, and improves the quality of content analysis. Furthermore, readers of studies using content analysis gain deeper insights into the quality of the data. They can form an opinion regarding which parts of the data correctly reflect a phenomenon, and where the data may be biased. A very helpful tool for evaluating the quality is the assignment-error matrix which provides information on how the categories may confound one another.

Based on the results of this study, the authors of this paper recommend complementing the data of learning analytics by using the textual data of students. This approach offers the opportunity to gain deeper insights into the cognition of learners while building a bridge to the conceptual work of different scientific and vocational disciplines. Furthermore, AI applications should present the reliability of every single category by using the new Iota Concept. It appears reasonable that the content analysis used in scientific studies should report the reliability of every single category using the cut-off values presented in **Table 9**. We recommend using the minimum iota, as this value ensures a minimal reliability standard for every single category that cannot be compensated by the superior reliability of other categories.

The calculation can be easily done with the package *iotarelr* which was developed simultaneously to this paper. Currently the package is only available at github. A submission to CRAN is planned in the future. News, introductions, and guides on how to use the package can be found via the project page[2].

Regarding the configuration of AI, the results in **Table 9** show which hyperparameters should be explicitly configured, and which parameters should be minimized/maximized. Of particular importance are the filter method and the number of features/words used for creating AI, since the standardized coefficients are relatively large. The aim of training AI under the condition of small sample sizes is the creation of a compressed textual representation relying on the most important information. Based on this study, JMIM can be used for selecting relevant words. The number of words should then be clearly filtered to about 5% of the initial number or even lower. Further research could focus the impact of other methods to create compressed textual representations. Technically, factor analysis, latent semantic analysis, latent Dirichlet allocation, and global vectors may be interesting for this purpose.

LIMITATIONS AND FURTHER RESEARCH

The limitations of this study point toward the need for future research. First, the simulation study uses only a simple linear relationship for ordinal data to derive cut-off values for the new measures. Further studies could investigate more complex relationships for ordinal and nominal variables. Second, the dependent variable is assumed as being measured with perfect reliability. This assumption does not hold in practice. Consequently, the cut-off values have to be higher. To derive more meaningful cut-off values, further simulation studies should therefore vary the reliability of the dependent variable. Third, the simulation study assumes that the true reliability is the same for all categories. Further research should investigate the relationship between iota and the true reliability for more varying values between the categories. Forth, the data for training AI was gathered from existing studies. The structure of the data did not allow to include an indicator of the quality of the initial data into the analysis although the study by Song et al. (2020) showed that this is a critical factor. Therefore, future studies should include corresponding indicators in their analysis.

In the current study, only a limited number of filter methods and kinds of AI could be applied. Additional research should include more of these different methods to find the best algorithms for varying conditions.

[2]https://fberding.github.io/iotarelr/

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication. The iota concept itself was developed by FB.

FUNDING

These simulations were performed at the HPC Cluster CARL, located at the University of Oldenburg (Germany), and funded by the DFG through its Major Research Instrumentation Program (INST 184/157-1 FUGG), and the Ministry of Science and Culture (MWK) of the State of Lower Saxony in Germany.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.818365/full#supplementary-material

REFERENCES

Alpaydin, E. (2019). *Maschinelles Lernen*. Berlin: DE GRUYTER.

Alpizar, D., Adesope, O. O., and Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educ. Technol. Res. Dev.* 68, 2095–2119. doi: 10.1007/s11423-020-09748-7

Anders, Y., Kunter, M., Brunner, M., Krauss, S., and Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychol. Erzieh. Unterr.* 57, 175–193. doi: 10.2378/peu2010.art13d

Baker, F. B., and Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Cham: Springer International Publishing.

Bennasar, M., Hicks, Y., and Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Syst. Appl.* 42, 8520–8532. doi: 10.1016/j.eswa.2015.07.007

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., et al. (2018). quanteda: an R package for the quantitative analysis of textual data. *J Open Source Softw.* 3, 1–4. doi: 10.21105/joss.00774

Berding, F. (2019). *Rechnungswesenunterricht: Grundvorstellungen und ihre Diagnose*. Augsburg: Hampp.

Berding, F., and Jahncke, H. (2020). "Die Rolle von Grundvorstellungen in Lehr-Lern-Prozessen im Rechnungswesenunterricht – Eine Mehr-Ebenen-Analyse zu den Überzeugungen von Lehrkräften und Grundvorstellungen, Motivation, Modellierungsfähigkeit und Noten von Lernenden," in *Moderner Rechnungswesenunterricht 2020: Status quo und Entwicklungen aus wissenschaftlicher und praktischer Perspektive*, eds F. Berding, H. Jahncke, and A. Slopinski (Wiesbaden: Springer), 227–258. doi: 10.1007/978-3-658-31146-9_11

Berding, F., Riebenbauer, E., Stütz, S., Jahncke, H., Slopinski, A., and Rebmann, K. (2022). Performance and Configuration of Artificial Intelligence in Business Education Learning Analytics Applications. A Content Analysis-Based Approach. *Preprint* doi: 10.31235/osf.io/trvcy

Berding, F., Stütz, S., Jahncke, H., Holt, K., Deters, C., and Schnieders, M.-T. (2021). Kosten und leistungen, eigenkapital und fremdkapital. Grundvorstellungen von realschülerinnen und realschülern sowie studierenden und ihr einfluss auf lernprozesse und lernerfolge. *Z. Berufs Wirtschaftspädagog.* 117, 560–629. doi: 10.25162/zbw-2021-0023

Bergstra, J., and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* 13, 281–305.

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., and Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: from the general to the applied. *J. Comput. High. Educ.* 26, 87–122. doi: 10.1007/s12528-013-9077-3

Bloom, B. S. (1984). The 2 Sigma Problem: the Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educ. Res.* 13:4. doi: 10.2307/1175554

Bonifay, W. (2020). *Multidimensional item response theory*. Los Angeles: SAGE.

Borges Völker, E., Wendt, M., Hennig, F., and Köhn, A. (2019). "HDT-UD: A very large Universal Dependencies Treebank for German," in *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, eds A. Rademaker and F. Tyers (Stroudsburg, PA, USA: Association for Computational Linguistics), 46–57.

Brom, C., Stárková, T., and D'Mello, S. K. (2018). How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educ. Res. Rev.* 25, 100–119. doi: 10.1016/j.edurev.2018.09.004

Cerasoli, C. P., Nicklin, J. M., and Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: a 40-year meta-analysis. *Psychol. Bull.* 140, 980–1008. doi: 10.1037/a0035661

Cohen, J. (1988). *Statistical Powe Analysis for the Behavioral Sciences*. New York: Taylor & Francis.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, London: The Guilford Press.

ElSayed, A. A., Caeiro-Rodríguez, M., Mikic-Fonte, F. A., and Llamas-Nistal, M. (2019). "Research in Learning Analytics and Educational Data Mining to Measure Self-Regulated Learning: a Systematic Review," in *The 18th World Conference on Mobile and Contextual Learning* (Netherlands: Delft University of Technology).

Euler, D., and Hahn, A. (2014). *Wirtschaftsdidaktik*. Berne Bern: Haupt Verlag.

Feng, G. C., and Zhao, X. (2016). Do Not Force Agreement. *Methodology* 12, 145–148. doi: 10.1027/1614-2241/a000120

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* 73, 220–239. doi: 10.1016/j.eswa.2016.12.035

Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. *Int. J. Res. Mark.* 36, 20–38. doi: 10.1016/j.ijresmar.2018.09.009

Hayes, A. F., and Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.* 1, 77–89. doi: 10.1080/19312450709336664

Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Kallmeyer.

Hove, D., ten, Jorgensen, T. D., and van der Ark, L. A. (2018). "On the Usefulness of Interrater Reliability Coefficients," in *Quantitative Psychology*, eds M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer International Publishing), 67–75. doi: 10.1007/978-3-319-77249-3_6

Ifenthaler, D., and Widanapathirana, C. (2014). Development and Validation of a Learning Analytics Framework: two Case Studies Using Support Vector Machines. *Technol. Knowl. Learn.* 19, 221–240. doi: 10.1007/s10758-014-9226-4

Jaakonmäki, R., vom Brocke, J., Dietze, S., Drachsler, H., Fortenbacher, A., Helbig, R., et al. (2020). *Learning Analytics Cookbook*. Cham: Springer.

Jahncke, H. (2019). *Selbst-)Reflexionsfähigkeit: Modellierung, Differenzierung und Beförderung mittels eines Kompetenzentwicklungsportfolios*. München: Hampp.

Karst, K., Schoreit, E., and Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern. *Z. für Pädagog. Psychol.* 28, 237–248. doi: 10.1024/1010-0652/a000133

Kleesiek, J., Murray, J. M., Strack, C., Kaissis, G., and Braren, R. (2020). Wie funktioniert maschinelles Lernen? *Der Radiologe* 60, 24–31. doi: 10.1007/s00117-019-00616-x

Krippendorff, K. (2016). Misunderstanding Reliability. *Methodology* 12, 139–144. doi: 10.1027/1614-2241/a000119

Krippendorff, K. (2019). *Content Analysis: An Introducation to Its Methodology*. Los Angeles: SAGE.

Kühne, V. (2021). *Modellierungskompetenz im Rechnungswesenunterricht: Eine empirische Analyse von Schulbüchern*. Master thesis. Oldenburg.

Kulik, C.-L., and Kulik, J. A. (1991). Effectiveness of computer-based instruction: an updated analysis. *Comput. Hum. Behav.* 7, 75–97. doi: 10.1016/0747-5632(91)90030-5

Kursa, M. B. (2021). *praznik: Tools for Information-Based Feature Selection. Version 7.0.0*.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., et al. (2019). mlr3: a modern object-oriented machine learning framework in R. *J. Open Source Softw.* 4:1903. doi: 10.21105/joss.01903

Lanquillon, C. (2019). "Grundzüge des maschinellen Lernens," in *Blockchain und maschinelles Lernen: Wie das maschinelle Lernen und die Distributed-Ledger-Technologie voneinander profitieren*, eds S. Schacht and C. Lanquillon (Heidelberg: Springer), 89–142.

Larusson, J. A., and White, B. (2014). "Introduction," in *Learning Analytics: From Research to Practice*, eds J. A. Larusson and B. White (New York, NY: Springer New York), 1–12. doi: 10.1093/oso/9780198854913.003.0001

Liu, M.-C., Yu, C.-H., Wu, J., Liu, A.-C., and Chen, H.-M. (2018). Applying Learning Analytics to Deconstruct User Engagement by Using Log Data of MOOCs. *J. Inf. Sci. Eng.* 34, 1174–1186. doi: 10.6688/JISE.201809_34(5).0004

Long, D., and Magerko, B. (2020). "What is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA: ACM).

Lorena, A. C., Jacintho, L. F., Siqueira, M. F., de Giovanni, R., Lohmann, L. G., Carvalho, A. C., et al. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Syst. Appl.* 38, 5268–5275. doi: 10.1016/j.eswa.2010.10.031

Lovejoy, J., Watson, B. R., Lacy, S., and Riffe, D. (2016). Three Decades of Reliability in Communication Content Analyses. *Journal. Mass Commun. Q.* 93, 1135–1159. doi: 10.1177/1077699016644558

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., et al. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. *Front. Psychol.* 11:580820. doi: 10.3389/fpsyg.2020.580820

Mayer, R. E. (2019). "Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principle," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (New York: Cambridge University Press), 345–368. doi: 10.1017/cbo9781139547369.017

Mayer, R. E., and Fiorella, L. (2019). "Principles for reducing extraneous processing in multimedia learning: Coher-ence, signaling, redundancy, spatial contiguity, and temporal contiguity principle," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (New York: Cambridge University Press), 279–315. doi: 10.1017/cbo9781139547369.015

Mayer, R. E., and Pilegard, C. (2019). "Principles for managing essential processing in multimedia learning: Seg-menting, pre-training, and modality principle," in *The Cambridge handbook of multimedia learning*, ed. R. E. Mayer (New York: Cambridge University Press), 316–344. doi: 10.1017/cbo9781139547369.016

Means, B., Toyama, Y., Murphy, R., Bakia, M., and Jones, K. (2010). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Available Online at: www.ed.gov/about/offices/list/opepd/ppss/reports.html (accessed July 30, 2020).

Paek, I., and Cole, K. (2020). *Using R for item response theory model applications*. Abingdon, Oxon, New York, NY: Routledge.

Papilloud, C., and Hinneburg, A. (2018). *Qualitative Textanalyse mit Topic-Modellen: Eine Einführung für Sozialwissenschaftler*. Wiesbaden: Springer.

Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: importance of Hyperparameters of Machine Learning Algorithms. *J. Mach. Learn. Res.* 20, 1–32. doi: 10.1007/978-1-4842-6579-6_1

Reich, J. (2015). Education research. Rebooting MOOC research. *Science* 347, 34–35. doi: 10.1126/science.1261627

Riebenbauer, E. (2021). *Kompetenzentwicklung im Masterstudium Wirtschaftspädagogik. Längsschnittstudie zur Unterrichtsplanung im Rechnungswesen*. Bielefeld: wbv Media. doi: 10.3278/9783763970216

Richter, S. (2019). *Statistisches und maschinelles Lernen*. Berlin: Springer.

Rienties, B., Køhler Simonsen, H., and Herodotou, C. (2020). Defining the Boundaries Between Artificial Intelligence in Education, Computer-Supported Collaborative Lea128rning, Educational Data Mining, and Learning Analytics: a Need for Coherence. *Front. Educ.* 5:128. doi: 10.3389/feduc.2020.00128

Ryan, R. M., and Deci, E. L. (2012). "Motivation, personality, and development within embedded social contexts: An overview of Self-Determination Theory," in *The Oxford handbook of human motivation*, ed. R. M. Ryan (Oxford: Oxford University Press), 84–108. doi: 10.1093/oxfordhb/9780195399820.013.0006

Saura, J. R., Ribeiro-Soriano, D., and Zegarra Saldaña, P. (2022). Exploring the challenges of remote work on Twitter users' sentiments: from digital technology development to a post-pandemic era. *J. Bus. Res.* 142, 242–254. doi: 10.1016/j.jbusres.2021.12.052

Schneider, S., Beege, M., Nebel, S., and Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educ. Res. Rev.* 23, 1–24. doi: 10.1016/j.edurev.2017.11.001

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Ph.D. thesis. Frankfurt am Main: Universität Heidelberg.

Schreier, M. (2012). *Qualitative Content Analysis in Practice*. Los Angeles: SAGE.

Seufert, S., Guggemos, J., and Ifenthaler, D. (2021). "Zukunft der Arbeit mit intelligenten Maschinen: Implikationen der Künstlichen Intelligenz für die Berufsbildung," in *Künstliche Intelligenz in der beruflichen Bildung: Zukunft der Arbeit und Bildung mit intelligenten Maschinen?*, eds S. Seufert, J. Guggemos, D. Ifenthaler, H. Ertl, and J. Seifried (Stuttgart: Franz Steiner Verlag), 9–27.

Siemens, G., Dawson, S., and Lynch, G. (2013). *Improving the Quality and Productivity of the Higher Education Sector: Policy and Strategy for Systems-Level Deployment of Learning Analytics*. Available Online at: https://solaresearch.org/wp-content/uploads/2017/06/SoLAR_Report_2014.pdf (accessed Oct 31, 2020).

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., et al. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Polit. Commun.* 37, 550–572. doi: 10.1080/10584609.2020.1723752

Straka, M., and Straková, J. (2017). *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. Available Online at: https://www.aclweb.org/anthology/K17-3009.pdf (accessed July 03, 2020).

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., and Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: a second-order meta-analysis and validation study. *Rev. Educ. Res.* 81, 4–28. doi: 10.3102/0034654310393361

Therneau, T., Atkinson, B., and Ripley, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. Available Online at: https://CRAN.R-project.org/package=rpart (accessed May 9, 2022).

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* 46, 197–221. doi: 10.1080/00461520.2011.611369

Venables, W. N., and Ripley, B. D. (2007). *Modern applied statistics with S*. New York, NY: Springer.

Wang, J., and Wang, X. (2020). *Structural Equation Modeling: Applications Using Mplus*. Hoboken: Wiley.

Weinert, F. E., and Schrader, F.-W. (1986). "Diagnose des Lehrers als Diagnostiker," in *Schülergerechte Diagnose*, eds H. Petillon, J. E. Wagner, and B. Wolf (Weinheim: Beltz), 11–29. doi: 10.1007/978-3-322-87640-9_2

Wijffels, J., Straka, M., and Straková, J. (2019). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. Available Online at: https://CRAN.R-project.org/package=udpipe (accessed May 9, 2022).

Wong, J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., et al. (2019). "Educational Theories and Learning Analytics: From Data to Knowledge," in *Utilizing Learning Analytics to Support Study Success*, eds D. Ifenthaler, D.-K. Mah, and J. Y.-K. Yau (Cham: Springer International Publishing), 3–25. doi: 10.1007/978-3-319-64792-0_1

Wright, M. N., and Ziegler, A. (2017). ranger : a Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Soft.* 77, 1–17. doi: 10.18637/jss.v077.i01

Zhao, X., Feng, G. C., Liu, J. S., and Deng, K. (2018). We agreed to measur o measure agreement - Redefining r eement - Redefining reliability de-justifies eliability de-justifies Krippendorff's alpha. *China Media Res.* 14, 1–15.

Zhao, X., Liu, J. S., and Deng, K. (2013). Assumptions behind Intercoder Reliability Indices. *Ann. Int. Commun. Assoc.* 36, 419–480. doi: 10.1080/23808985.2013.11679142

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/feduc.2022.913530

Learn to Machine Learn *via* Games in the Classroom

Marvin Zammit, Iro Voulgari, Antonios Liapis and Georgios N. Yannakakis*

Institute of Digital Games, University of Malta, Msida, Malta

Artificial Intelligence (AI) and Machine Learning (ML) algorithms are increasingly being adopted to create and filter online digital content viewed by audiences from diverse demographics. From an early age, children grow into habitual use of online services but are usually unaware of how such algorithms operate, or even of their presence. Design decisions and biases inherent in the ML algorithms or in the datasets they are trained on shape the everyday digital lives of present and future generations. It is therefore important to disseminate a general understanding of AI and ML, and the ethical concerns associated with their use. As a response, the digital game *ArtBot* was designed and developed to teach fundamental principles about AI and ML, and to promote critical thinking about their functionality and shortcomings in everyday digital life. The game is intended as a learning tool in primary and secondary school classrooms. To assess the effectiveness of the ArtBot game as a learning experience we collected data from over 2,000 players across different platforms focusing on the degree of usage, interface efficiency, learners' performance and user experience. The quantitative usage data collected within the game was complemented by over 160 survey responses from teachers and students during early pilots of ArtBot. The evaluation analysis performed in this paper gauges the usability and usefulness of the game, and identifies areas of the game design which need improvement.

Keywords: machine learning, artificial intelligence, serious games, educational games, game analytics, digital literacy, supervised learning, reinforcement learning

OPEN ACCESS

Edited by:
Iza Marfisi-Schottman, EA4023 Laboratoire d'Informatique de l'Université du Mans (LIUM), France

Reviewed by:
Samia Bachir, Le Mans Université, France
Michael Kickmeier-Rust, University of Teacher Education St. Gallen, Switzerland

****Correspondence:***
Marvin Zammit
marvin.zammit@um.edu.mt

Specialty section:
This article was submitted to Digital Learning Innovations, a section of the journal Frontiers in Education

Received: *05 April 2022*
Accepted: *20 May 2022*
Published: *17 June 2022*

Citation:
Zammit M, Voulgari I, Liapis A and Yannakakis GN (2022) Learn to Machine Learn via Games in the Classroom. Front. Educ. 7:913530. doi: 10.3389/feduc.2022.913530

1. INTRODUCTION

Through the use of the world-wide web and access to applications such as social media, games, Google search, YouTube, and the Internet of Things, digital technologies are mediating children's learning, entertainment and social interactions, and are shaping their everyday lives (Rahwan et al., 2019). The online environment has become ubiquitous, and artificial intelligence (AI) and machine learning (ML) algorithms are employed across most digital platforms to track user preferences, suggest content, and even generate novel material for a wide variety of purposes (Yannakakis and Togelius, 2018). This pervasiveness can enhance the user experience but is fraught with controversy, as user data is often stored and possibly shared with third parties without clarity to the user. This may lead to breach of personal information, or more nefarious acts such as steering of public opinion or distortion of facts. Despite the implications of such digital technologies, children's conceptions of technology, its potential and implications may be vague, inaccurate or distorted (Druga et al., 2017; Mertala, 2019). It is therefore crucial—especially for the younger generations—to understand how AI is used, how it works, and what its pitfalls are. Critical thinking skills, such as making inferences, decision-making, and problem solving (Lai, 2011) have to be promoted. This is necessary for children to understand the implications and context of the online content they access, to recognize the use of AI and how it encroaches into their daily life and their social environment.

The work presented in this paper is situated in this context; we present ArtBot, a game that aims to provide young learners with the background knowledge of basic concepts behind AI and ML, and to show how basic algorithms can be used to solve different problems. In addition, we aim to highlight the challenges emerging from AI, thereby exposing students to the capabilities and limitations of these algorithms. Extensive research over the past few decades has shown the effectiveness and potential of digital games as learning tools; games may support motivation, engagement, and active participation of learners, enhance attention, involvement, and understanding of abstract and complex concepts (Hainey et al., 2016; Tsai and Tsai, 2020). Our game, ArtBot, builds upon this potential of games and aims to introduce young learners to the two fundamental processes of ML: supervised learning (SL) and reinforcement learning (RL). ArtBot was developed to be a teaching tool in classrooms to raise awareness and critical thinking about AI processes. The game was made available primarily through web browsers[1], but also for download on Windows operating systems, and as a mobile app on Android devices through the Google Play store. ArtBot comprises of two mini-games, one containing a level dedicated to SL, and the other consisting of ten levels dedicated to RL. The game narrative has the players tasked with retrieving art objects which have gone missing. The player is assisted by an AI helper (ArtBot) which they must train first to distinguish between statues and paintings, then to navigate rooms to collect statues scattered within them, avoid hazards and reach the exit. The game offers players customization of the avatar depicting ArtBot from a set of predefined models and color schemes.

ArtBot has been deployed on all platforms concurrently in April 2021, and included the collection of anonymous usage metrics. In our past publications (Voulgari et al., 2021; Zammit et al., 2021), we have focused on the educational design of ArtBot, how the requirements which emerged through focus groups with stakeholders were translated and adapted into a practical game design, and its implementation, deployment and initial reception. This paper, instead, explores the longitudinal usage of the ArtBot game in real-world settings over a period of almost 1 year.

Since its deployment to the general public, the game has been played by over 2,000 unique users across all platforms. This has supplied us with a substantial body of interaction data to analyse the game objectively, and draw conclusions about its design and interface based on user behavior. In this paper we review this data and try to obtain practical insights that expose the strengths and weaknesses of the game, and to evaluate whether the players' interaction with the game followed patterns intended by the design process.

2. GAMES FOR AI LITERACY

Using games for teaching AI is not an entirely new concept. Initiatives such as those of Clarke and Noriega (2003) and Hartness (2004), which involved a war simulation game and *Robocode*, respectively, introduced games for teaching AI algorithms to undergraduate computer science students. What has shifted, though, over the past few years is the framing of AI education through games and the age of the target group. Building upon the potential of games to support systems thinking, computational thinking, and understanding of complex concepts and processes (Clark et al., 2009; Voulgari, 2020), platforms, games, and applications to support AI literacy have been developed for learners as young as 4 years old, addressing the technical, societal, and ethical aspects of AI (Giannakos et al., 2020; Zammit et al., 2021).

Games, such as the commercial game *While True: Learn()*[2] and *ViPER* (Parker and Becker, 2014) are appropriate for younger students and aim to scaffold the players through understanding ML concepts such as optimization, loops, and model accuracy. In *While True: Learn()* players assume the role of a computer programmer who tries to develop a model for communicating with their cat and complete tasks using visual programming, while in *ViPER* players train a robot, through coding, to navigate its way on one of Jupiter's moons. Another commercial game, *Human Resource Machine*[3], which introduces the players to concepts such as automation and optimization by programming the employees in an office environment, has been adapted for students and teachers in educational contexts such as participation in the *Hour of Code*.[4]

Available platforms and tools allow learners to build their own ML models (engaging them in exploratory and constructivist learning practices) and to situate the models in the context of their own interests or real-world problems. *Google Teachable Machine* (GTM), for instance, has been used by 12–13 years old students, introducing them to ML concepts through the design of their own ML models, relevant to their interests or real-world problems (Toivonen et al., 2020; Vartiainen et al., 2021). Results have shown that GTM is appropriate for students with little or no programming experience. Situated in an appropriate learning context, activities through GTM allowed the students to exhibit design thinking, reason inductively about the quality of the datasets and the accuracy of their models, and show empathy to other people's needs in order to develop appropriate ML applications. Zimmermann-Niefield et al. (2019), using a tool they developed (*AlpacaML*), also engaged high school students in the design of their own ML models based on their own interest (i.e., athletic activities). The students experimented with the design of their models and reflected upon the characteristics of a good model and how the models work.

Approaches such as those of Turchi et al. (2019) and Microsoft's *Minecraft. Hour of Code: AI for Good*[5] also situate AI and ML into real-world problems. Turchi et al. (2019) used a combination of online and board game play to introduce AI concepts to students and professionals involving the protection of wildlife, while *Minecraft. Hour of Code: AI for Good* scaffolds

[1]Initially launched at http://learnml.eu/games.php and later moved to a dedicated website at http://art-bot.net/ on 22 March 2022.

[2]https://luden.io/wtl/

[3]https://tomorrowcorporation.com/humanresourcemachine

[4]https://tomorrowcorporation.com/human-resource-machine-hour-of-code-edition

[5]https://education.minecraft.net/hour-of-code

players to programme a robot to predict forest fires. Through such approaches learners may not only be introduced to concepts of AI and ML but also understand the role, potential, and impact of AI and ML applications in authentic and meaningful contexts.

Existing platforms, games, and applications seem to either provide an open-ended environment for learners to experiment, design and develop their own models applying concepts of AI and ML, or scaffold learners through a linear sequence of puzzles, to become familiar with AI and ML functions, processes, and algorithms (Voulgari et al., 2021). While the latter approach may facilitate novice students to understand basic principles of AI and ML algorithms, the open-ended approach allows students to engage in problem-solving tasks, reflect on their actions, assess and re-examine their progress and construct their knowledge by assuming a more active role, in line with constructivist and constructionist approaches (Kafai and Burke, 2015). In our game design, we tried to combine elements from both approaches; by guiding the players through linear tasks we aimed to introduce learners to core concepts of AI and ML, and through the more open-ended tasks, we provided space for experimentation, problem-solving, and reflection.

3. LEARNML AND ARTBOT

LearnML (or Learn to Machine Learn)[6] is a European-funded project aiming to develop digital literacy and awareness of AI usage in the digital landscape to learners who are exposed to these technologies from an early age. Its goal is to develop a toolkit for teachers and students which can be used primarily in a classroom environment, but also in non-formal learning settings (e.g., at home). The project involved the development of a number of educational games which teach different aspects of ML. These are supported by teaching materials that supplement the experience through classroom discussions to encourage reflection and critical thinking. A number of workshops and events for teachers and students were also organized in order to disseminate the work done and also gauge feedback directly from stakeholders (Voulgari et al., 2021). The impact on students, teachers, and their needs was always a priority during the development of ArtBot, which was part of the LearnML project and shared its broader goals. ArtBot was designed in collaboration with educators, with requirements collected through participatory design workshops held in three countries (Greece, Malta, and Norway) and included participation of all stakeholder categories: e.g., teachers, students, and AI researchers (Zammit et al., 2021).

The goals of ArtBot were distilled from the stakeholder needs and consisted of the following (Voulgari et al., 2021):

- To introduce the process of supervised learning, including terminology and concepts of training and testing datasets, classification, labeling, image recognition, decision trees, and prediction accuracy, and outline their role and behavior in an AI system.
- To introduce reinforcement learning and related concepts, such as rewards and penalties, learning rate, exploration, exploitation, and pathfinding.
- To show that the design decisions behind the implementation of an AI system have a considerable bearing on its behavior and outputs, thereby highlighting the fact that human bias may seep through the workings of the algorithms.
- To provoke reflection and discussions on the impact of AI systems in everyday life situations e.g., facial recognition, self-driving vehicles, etc.

With these objectives in mind, two mini-games were developed within ArtBot, focusing on supervised learning (SL) and reinforcement learning (RL), respectively. Specifically, regarding SL, concepts such as training set, testing set, data labeling, classification. and decision trees were introduced. For RL, we focused on introducing concepts such as rewards, learning rate, and exploration. The premise of the game was that a number of statues have gone missing, and the players are tasked with their retrieval. To assist in their quest, the players are given an autonomous helper (ArtBot). We tried to set the story in a meaningful narrative background since narrative seems to motivate the students, support understanding of abstract concepts and the construction of mental models, and re-frame the activities and challenges of the game into an authentic context (Glaser et al., 2009). ArtBot must first be trained to distinguish between statues and paintings, before it can be sent to retrieve the missing statues. Therefore, the SL mini-game is played first, and once completed the RL mini-game becomes available; in general, players need to complete the previous mini-game or level in order to proceed to the next one. Each mini-game is described below, while more details on the design, interfaces, and algorithms included in ArtBot are provided by Zammit et al. (2021).

3.1. Description of the Supervised Learning Mini-Game

Supervised learning (SL) is an umbrella term for ML algorithms which are used when a considerable amount of data pertaining to the problem is available (Zhou and Liu, 2021). SL is commonly applied to classification problems. An existing set of labeled data is processed through the algorithm, which tries to find some complex function that accommodates the majority of points. The available data is usually split into a training set, which is used to teach the model, and a testing set, which is used to verify the accuracy of the trained model on unseen data. Some inherent problems with SL are that data labeling is a laborious process, and that the data itself may be biased, or even incorrect, leading to this bias being learned by the model itself.

The SL mini-game tasks players to label a number of images of paintings and statues; after the labeling process is complete, ArtBot uses a decision tree based on the supplied labels to classify hitherto unseen images. All images were obtained from the Open Access Artworks[7] collection of the Metropolitan Museum of Art in New York, USA, and are photographs of real paintings and statues.

[6]http://learnml.eu/

[7]https://www.metmuseum.org

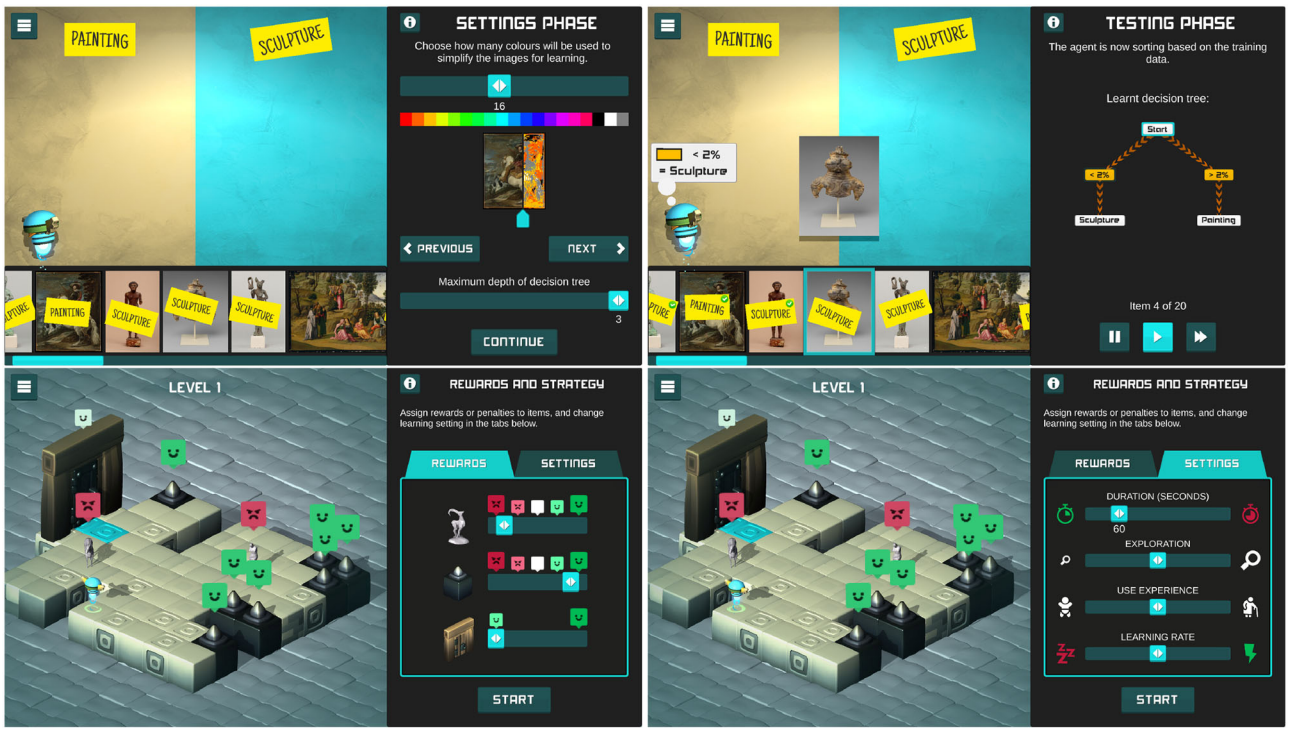


FIGURE 1 | Screenshots of ArtBot, showing (in order) the SL mini-game settings (top left) and the resulting classification (top right) screens. The RL mini-game offers two sets of settings to the players: rewards and penalties of the various objects (bottom left), and RL parameters (bottom right).

Players are allowed to assign incorrect labels, experiment, and see how this affects ArtBot's learned classification skills. The classification was simplified to a left swipe (to label an image as painting) or right swipe (to label it as statue) in order to speed up the process. However, this is an inherently repetitive task and can rapidly get boring. To mitigate this, we limited the classification task to 20 images, and also provided an auto-sort button which assigns the ground truth label to each image automatically.

In order to decrease the computational resources required to process and group the dataset images, the SL algorithm splits the pixel spectrum into a smaller number of colors which is controlled by the players. A decision tree is then trained on the training set that has been labeled by the player using the C4.5 algorithm (Quinlan, 1992), and used to classify the image as a painting or statue depending on the pixel count in each of the resulting color bins. The player is also given control on the maximum allowed depth of the decision tree. The interface for this mini-game is shown in **Figure 1**. Note that when showing the player the results of the SL mini-game, the training accuracy is shown as well as a testing accuracy on a set of 20 images that are unseen by both the players and the decision tree, and have been labeled correctly by the game's designers. This training and testing accuracy introduces learners to the concept of training and test sets, and illustrates how supervised learning can be used to predict patterns in unseen data but can suffer from overfitting to patterns in the training data.

3.2. Description of the Reinforcement Learning Mini-Game

Reinforcement Learning (RL) is applied when the problem at hand requires a policy or a behavior which will generate the expected solution (Sutton and Barto, 2018). In such cases, there is generally either no predefined dataset available, or the problem is incongruous with a structured one. The learning process starts with a random policy (i.e., taking actions at random) while rewards or penalties are awarded for each action. The algorithm updates its knowledge about the problem according to these rewards, over a repeated number of trials—or episodes. The process tries to maximize the rewards obtained by changing its policy and observe the resulting performance. Some of the issues with this approach are the lengthy "trial and error" approach, as well as the challenge of balancing out how much weight the algorithm should put toward its acquired knowledge vs. the exploration of yet unknown actions.

In the RL mini-game, the player oversees how the ArtBot can learn to navigate through 10 different levels, avoiding hazards (spikes), collecting statues, and finding the exit. Players can set rewards and penalties to be used in the learning algorithm for each type of game object. In addition they are given a number of controls over RL parameters, such as the learning rate, the balance between exploration of new areas and utilization of already discovered information, and the total time allowed for training. When players choose rewards and parameters, the game will use a basic Q-learning algorithm (Watkins and Dayan, 1992) to negotiate a path across the level. Initially ArtBot will start

taking random steps in the environment and record the resulting rewards. As more rewards are encountered across the different episodes, the agent learns an optimal path that maximizes the total reward.

The players can progress to the next level if the exit is found and at least one statue is collected. They can however change the settings and retry each level at will. The user interface for this mini-game and the corresponding settings available to the players are shown in **Figure 1**.

4. DEPLOYMENT AND DATA COLLECTION

ArtBot was developed in the Unity game engine[8] due to the engine's capabilities to deploy to multiple platforms. We made use of Unity's integrated analytics service to collect in-game events anonymously across all platforms in order to better understand the performance of the game in terms of adoption, use, and player experience. ArtBot was deployed on the 8th April 2021 on the LearnML website[9]; however, it was shown to some focus groups even earlier as discussed in Section 4.2. The game can be played directly inside web browsers supporting WebGL technology, downloaded for Microsoft Windows operating systems, or downloaded to Android devices from the Google Play Store.[10] The game has been localized in English, Greek, and Norwegian languages, to facilitate a more widespread adoption in the countries of the LearnML project partners.

4.1. Collection of Game Analytics

The data presented in this document covers a period from the launch date (8 April 2021) up to 28 February 2022, although data collection is still ongoing.

Figure 2 shows the total number of distinct users as well as the monthly active users. Player uniqueness can be ascertained for the Windows and Android platforms, but web browser sessions are less identifiable (e.g., anonymous browsing modes in the web browser client will make different sessions by the same user appear as different users). However, we assume that this measurement error is not significant enough, especially since there is no apparent benefit for the users to obfuscate their activities in the game. **Table 1** shows the distribution of users across the platforms over which the game was deployed. It is interesting to note that the majority of the players chose to play the game through the browser, which indicates that facilitating immediate launch of the game (as opposed to the download and installation required on the other platforms) is important to the end-user. This is particularly relevant as the game is (also) intended for classroom use, and special privileges are usually required on public computers for installing new software; instead, the browser experience is available to all. It is also worth mentioning that most players that launched the app also initiated a game, and played at least the first SL mini-game.

[8]https://unity.com

[9]http://learnml.eu/games.php

[10]https://play.google.com/store/apps/details?id=com.InstituteofDigitalGames.ArtBot

The language change within the game was also monitored, and it was noted that 148 users (i.e., in 6.7% of game starts) switched to Greek, but no user has so far selected Norwegian. The game was launched in English and Greek, then updated with Norwegian localization on 2 September 2021, which has some bearing on this resulting lack of adoption.

4.2. Collection of User Evaluations *via* Surveys

Beyond anonymous data collected from log files, an online survey was also used to examine players' attitudes toward the game. Two surveys were designed: one for students (130 participants) and one for educators (35 participants). Seventy-four percent of the educators identified as female and 26% as male, while for the students 51% identified as male, 45% as female, and 4% preferred not to answer. Mean (M) age of the students was 14.3 years old, with a standard deviation (SD) of 5. The surveys included open questions regarding the positive and negative aspects of the game, and closed questions (5-item Likert scale) on whether they enjoyed the game and its learning potential. Gaming frequency varied among students and educators, reporting from 0 to 30 h of game playing over the past week for students ($M = 7.6$, SD $= 10.6$) and from 0 to 10 h for educators ($M = 1.6$, SD $= 2.6$). The educators came from a wide range of fields such as physics, language, mathematics, and information technology.

ArtBot was disseminated through the following events and avenues, where the game was demonstrated to participants, the participants were asked to play the game for a few minutes, and then asked to complete an evaluation survey:

- Distributed to secondary education teachers and students of a private school in Athens, Greece, in March 2021 (before the official launch).
- Demonstrated during online workshops with primary and secondary education teachers, mainly in computer science but also teachers from science education, linguistics and arts, in the framework of the Athens Science Festival in March 2021.
- Demonstrated during an online seminar mainly for secondary education teachers, organized by the 3rd Secondary Education Office in Attica, Greece, in May 2021.
- Demonstrated as part of an online teacher training event, in June 2021, organized in Malta. Participants were primary and secondary education teachers from a wide range of fields such as computer science, mathematics, economics, biology, Maltese, and ethics/religion, as well as other stakeholders (e.g., heads of school, school inspectors, researchers) from state, private, and church schools.
- Showcased and tested during a 3-day teacher training event for primary and secondary education teachers in October 2021.
- Showcased during a LearnML Info Day event in October 2021, mainly addressing educators and researchers.
- Shown as part of a keynote speech at the 3rd International Conference on Digital Culture & AudioVisual Challenges, addressing researchers and lecturers from a wide range of academic fields, in May 2021.

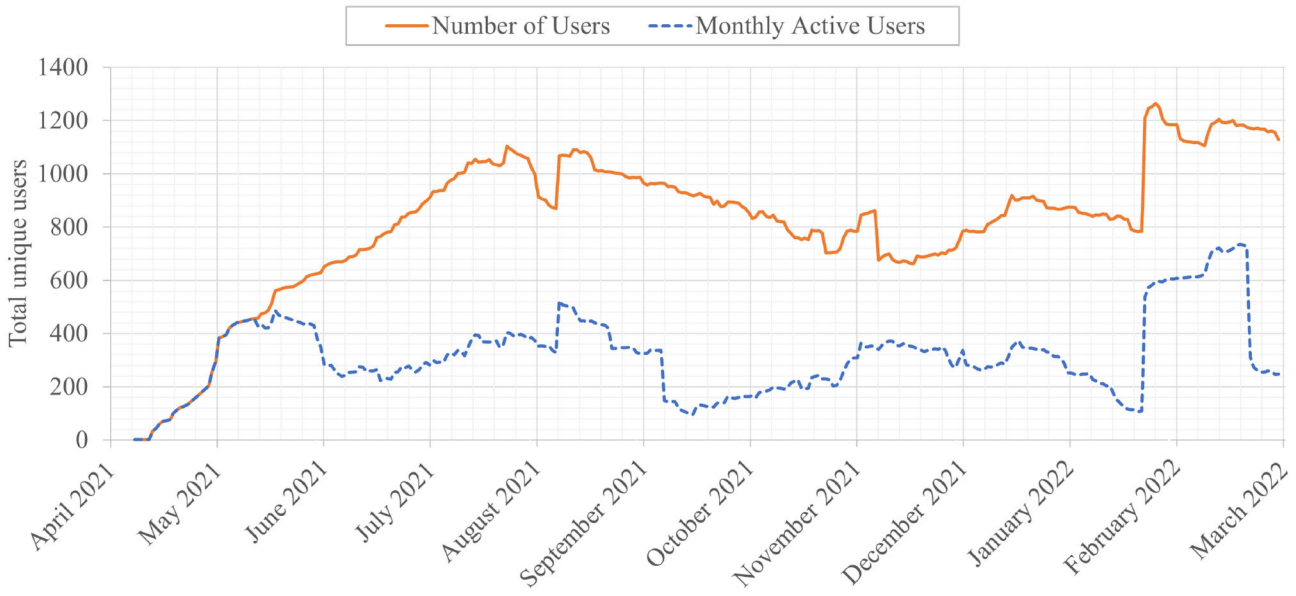


FIGURE 2 | Usage statistics showing the total number of unique users and the monthly active users over the data collection period.

TABLE 1 | Unique players sorted by platform.

	Total users	Browser	Android	Windows
Unique user visits	2,484	2,265	154	65
Unique user game starts	2,222	2,034	131	57

5. DATA ANALYSIS

Since its first launch in April 2021, the ArtBot game has been played by a total of 2,222 unique users. The users' interaction data with the game in general, and its two constituent mini-games around supervised learning (SL mini-game) and reinforcement learning (RL mini-game) are analyzed below, while the feedback of users (students and educators) to surveys solicited during dedicated events (see Section 4.2) is analyzed in Section 5.4.

5.1. General Use

As a first indication of the engagement of players with the game, we explore how many unique users interacted with different portions of the game. Since the game progresses sequentially from the menu page to the SL mini-game and then to the RL mini-game (the latter consisting of 10 levels), we observe how many users visited each portion in **Figure 3**. The SL game and the first RL level had a large number of players and a relatively high completion rate (82 and 74%, respectively), but the number of players moving on to later levels of the RL game decreases drastically after that. The completion rate remained high, indicating that persevering players were still engaged with the game. However, the high drop in the amount of players might indicate that the different levels of RL mini-game did not offer enough novelty to secure player retention.

Another practical metric for user engagement is the duration of playtime in each level. This proved to be a challenging measure to evaluate for the online playable version of the game. We noticed a number of users with extremely long times spent in each level (e.g., over an hour), which indicates that the game was most probably left running unattended in the browser while the user switched to another activity. Since the browser version was the most popular platform being used (see **Table 1**), this practice introduced a number of outlying data which skewed the statistics. The overall regular usage, however, was frequent enough to mitigate this, and we manually removed the outliers in terms of duration for this analysis.

When considering each level per mini-game, the mean time spent by the player in each of the levels was 4.2 min, although there was considerable variation. The SL mini-game and each level of the RL mini-game had a similar duration (3.8 min on average for SL, 4.4 for each RL level).

At every portion of the game, information screens were made available for additional explanation of how to use the game controls as well as further clarification about the underlying ML process. This information changed at every phase of the mini-game, updating the information with more relevant instructions and facts related to that specific part. Since the information was verbose, the display of these screens was entirely optional, and was displayed at the request of the player. Very few players opened these informational screens; the data indicates that most players who opened the information screen only did so during the first phase of each mini-game. This could be an indication that it was not clear to them that the content of that screen changed during subsequent phases. That said, the help options in the RL mini-game were requested more than two times as often as the respective help options for the SL mini-game. This indicates that

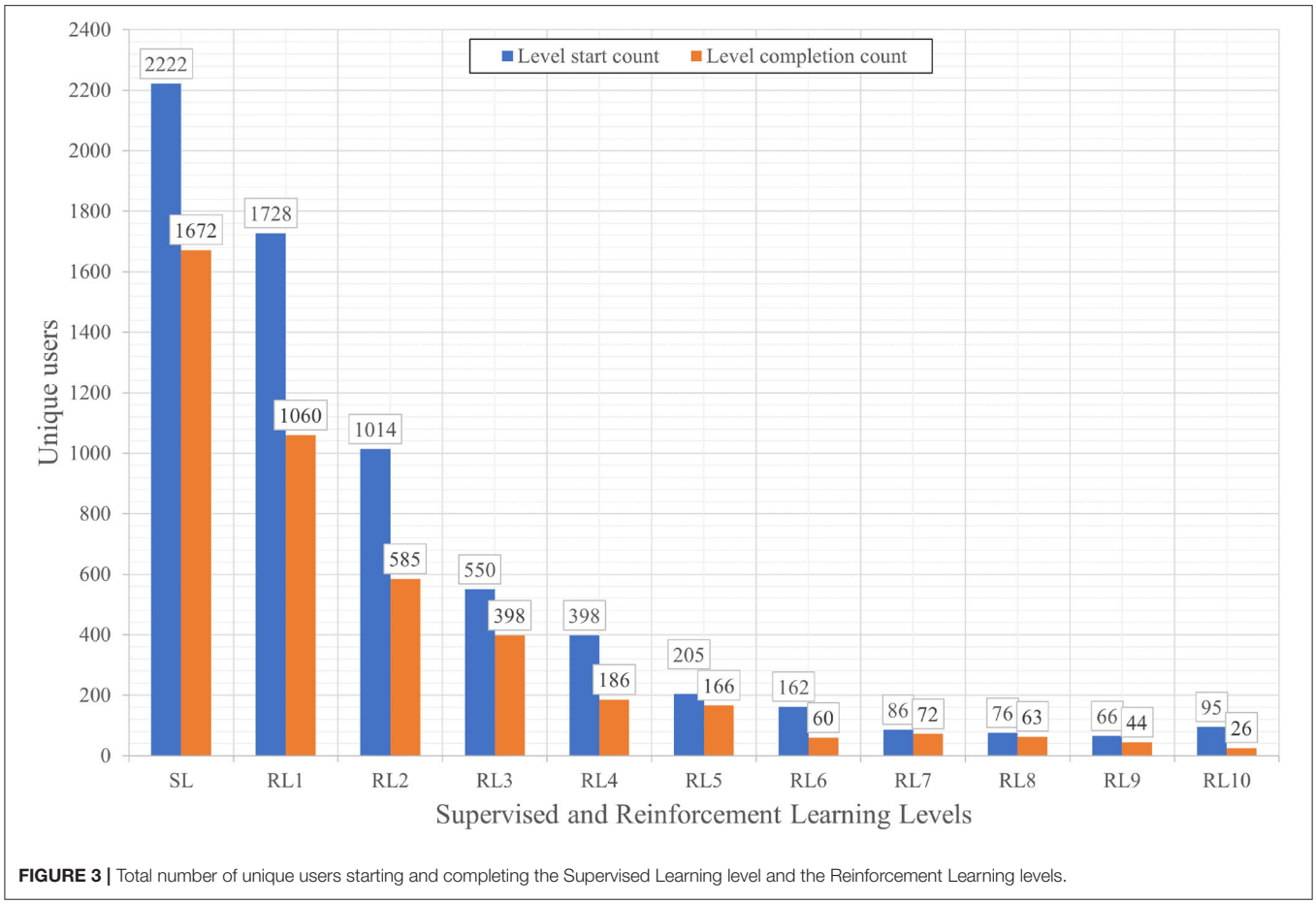


FIGURE 3 | Total number of unique users starting and completing the Supervised Learning level and the Reinforcement Learning levels.

the many parameters that could be tweaked in the RL sessions required more explanation than the few and intuitive options for the SL sessions.

At the beginning of the game the players are given the option to customize their avatar, selecting between five different avatars for ArtBot and three different color schemes per avatar. During the requirements collection stage of the game development process, it was determined that there exists a widespread misconception that AI is used mostly in robots. Consequently, the avatars were intentionally created without any anthropomorphic or highly technological connotations which may misconstrue them as being robots (Zammit et al., 2021). The avatar choices are shown in **Figure 4**. For ease of reference throughout this text the avatars are referred to as A1 to A5, each having three possible color schemes. The default avatar (A1 with the first color scheme) was the most commonly selected (36% of users). However, customization did occur and all possible avatar and color variations were chosen by at least 16 users. The most popular avatars (across all color schemes) were in order: A1 (54% of users), A2 (19%), A4 (11%), A3 (9%), A5 (7%). The more abstract avatars (A3, A5) were less commonly picked, while more anthropomorphic avatars (with bilateral symmetry and a distinction between front and back) were preferred despite our efforts to avoid these connotations.

5.2. The Supervised Learning Mini-Game

Due to our concern about the repetitiveness of the manual labeling process in the SL mini-game, we monitored the usage rate of the auto-sort function. Auto-sort was used to label the images automatically in 65% of the games played. Following the completion of an SL mini-game session, 25% of players repeated the labeling process, and 32% of players opted to change the parameters of the algorithm to see how the learned classification changed. These findings indicate that although some players were interested enough in the process to test different settings, the manual labeling activity was either not enjoyable, or does not yield enough of a different outcome for the experience to be considered worth the player's while.

Another important metric is *when* players were choosing to retry the level by changing the settings or relabeling the images. We noted each player's retries and what the resulting accuracy of the testing set was before and after the retry action: 32% of retries resulted in a greater accuracy, 28% regressed to a worse accuracy and the remaining 40% showed no change in accuracy. This suggests that the players were trying to improve their accuracy score, but the high rate of unvarying accuracy also implies that the settings we offer in the game might not allow enough room for improvement.

FIGURE 4 | The avatars which players can choose to represent ArtBot. In this paper they are referred to as A1 to A5 in sequential order from left to right.

The accuracy of labeling by the players compared to the ground truth was usually above 90%, even when ignoring the high number of auto-sorted runs; almost 99% of users that did manual labeling had 90% or above labeling accuracy on the training set. A few users labeled the images (paintings vs. statues) with lower accuracy, among which 10 unique users classified all paintings as statues and vice versa. Since the classification task itself is very easy for a human being, we deduce that this mislabeling was intentional and exploratory, which was among our initial objectives. Mislabeling the dataset would be a good entry point for discussion about the accuracy of trained models and the role of AI developers and data quality. However, players did not explore such disruptive labeling strategies which indicates the need to add prompts for the players to try to mislabel the data and reflect upon the results. Moreover, since the vast majority of users classified all images correctly (73% of players who did manual labeling), this would indicate that the task is trivial, and why auto-sorting was used to avoid it.

When considering the accuracy of the AI at the end of the supervised learning process, data shows that accuracy on the training dataset was very high, as expected (≥90% accuracy in 99% of sessions). The accuracy on a hitherto unseen testing dataset however, was much lower on average (67% across all sessions) and varied much more wildly. In 93% of the games played, the test accuracy was between 50% and 90%, and only went higher than that in 4% of sessions. Although it is normal for accuracy to drop during inference on a testing dataset, the reason why it is so pronounced here is probably the small amount (20) of labeled images that are used to train the algorithm. We opted for this small number to strike a good balance between accuracy and the tedium of the manual labeling task if extended to larger training datasets. This shortcoming, however, could act as a trigger for discussion with the learners on the factors affecting the quality of the trained model. A relevant note has actually been added at the information panel regarding the relation of the size of the training set and the accuracy of a model.

An analysis of the algorithmic settings used shows that the full range was used both for the color bin settings (from 4 to 32 in steps of 8) as well as for the maximum depth of the decision tree (1 to 3). The mean number of colors used as inputs to the SL algorithm was 15.0 across all players, with a standard deviation of 9.6. The high variance indicates that the color bin settings was being changed across replays. The tree depth was set to its maximum allowed value of 3 by 87% of the players, while 7.6% of players used a tree depth of 2 and 5.1% used a tree depth of 1. Overall, the large variance of algorithmic settings, accompanied by the variance in testing accuracy, points to some exploratory behavior from players in order to improve the behavior of the SL algorithm.

5.3. The Reinforcement Learning Mini-Game

The ArtBot mini-game focusing on reinforcement learning (RL) is richer in content, comprising of ten distinct levels when compared to the single SL mini-game. It also has a larger set of RL parameters for the player to explore, and its visuals are more congruent with those of commercial games. It was thus foreseeable that the engagement time for this part would be higher, as discussed in Section 5.1.

This mini-game includes a training time for the AI agent (between 30 s and 3 min) which can be chosen by the user. The mean training time set by players varied between 70 and 84 s across the different levels, with later levels (RL7 to RL10) falling closer to the upper end of this range. Since the game only allows training times to be in increments of 30 s, most players chose shorter training times (60 or 90 s).

Since rewards drive the AI behavior in RL problems, we observe what rewards and penalties the players assign to different game objects across the ten levels. Players may assign a reward of values between 1 and 5 to exits, and between -5 and 5 to hazards and collectibles. We monitored the mean and standard deviation of the assigned rewards across the different levels, which yielded useful insight. The standard deviation was consistently high in earlier levels of the RL mini-game (RL1 to RL5), indicating that players were assigning different rewards and penalties to explore how the learning of the agent is affected. In later levels (RL6 to RL10) the variance drops, indicating that those players who made it that far had developed some intuition for the more optimal assignments of these values. The mean reward for reaching the exit was consistently between 3.5 and 3.9 across levels. Players were often not assigning the maximum reward for the exit, in order to allow the agent to explore and obtain rewards from collectibles while learning. The mean reward for reaching collectibles was around 3.5 for the initial levels, then increased to around 4.3 at later levels. This again confirms more frequent variations in the settings in the earlier levels than in later levels, and that collectibles were prioritized over exits in order to promote collection of more statues in each level. The mean rewards for reaching hazards followed a reverse pattern, with negative rewards of −3 at earlier levels, up to −4.2 later in

the game. It also indicates an understanding by the players of the relation between the rewards or penalties assigned and the behavior of the agent.

We noted a different behavior with respect to the algorithm parameters. The variation within each RL parameter across sessions did not change much from level to level. This is interesting, as we were expecting that players would eventually find the best parameter setup and keep it consistent at later levels. The mean learning rate was consistently high throughout ($>$0.9), whereas the exploration rate had the highest variance from level to level, ranging between 76 and 89%. It is reasonable to set a high learning rate when training time is limited. Since the players were trying to collect more treasures before reaching the exit, it also makes sense for the exploration rate to be an intuitive parameter to vary. This indicates that players understood the underlying principles of reinforcement learning, and how human-determined settings have the potential of varying the outcome of an AI algorithm.

We also logged the results of the final runs after training is complete, as they offer additional insights to the individual level design in addition to AI behavior (and players' AI tuning priorities). The exit was found most of the time (above 80% of the time in most levels), but two levels were noteworthy. After training, in level RL4 the agent could reach the exit 44% of the time, indicating that the agent found it difficult to reach the exit across all players' attempts. This level contains the largest number of hazards from all levels, and a long distance between the statues and the exit, which indicates that the level design was indeed more taxing. For level RL9, in contrast, the exit was discovered by the trained agents in all cases across players. This may be in part because players who kept playing the game for nine levels so far were only those dedicated and knowledgeable enough of the RL parameters to achieve such performance; in RL10 for instance the trained agent reached the exit 90% of the time, which is also a high completion rate.

Figure 5 showcases the differences in layout between these two levels, which seemingly had a strong impact on the behavior of the agent. For RL4, there are spikes in the direct path from ArtBot's starting position to the exit, leading to the low completion rate. For RL9, there is a clear path from ArtBot's starting position to one statue and then the exit, but the other statue is behind a number of spikes and far from the path to the exit, leading to the low collection ratio.

It follows that finding the exit in each level is relatively easy with one exception. On the other hand, the (optional) task of finding all collectibles in the level before reaching the exit is more challenging. Indeed, the ratio of collectibles reached in the best trained agent varied significantly between levels. The most difficult levels in which to accomplish this task were RL10, RL6, RL9 with collection ratios of 10, 16, and 21%, respectively. Interestingly, while RL9 and RL10 were played only by a few (presumably expert) players who managed high completion rates, they had some of the lowest collection ratios; this may point to fatigue from the part of the players regarding the optional task of finding all collectibles.

The game allows players to stop training, which is useful if the learning process does not appear to be productive. The data shows that players tended to use this function mostly in the first four levels, then the rate gradually decreased; two exceptions were RL6 and RL10, which also had a large number of stops. This is congruent to the previous finding that these two levels were the most difficult in terms of collecting statues, indicating that the users were noticing that only one statue was being collected and stopped the training to revise the settings. In a similar fashion, levels RL6, RL4, RL10 were the ones in which the players most frequently changed the settings and retried the learning after the training was allowed to complete. These results keep underlining the difficulty in reaching the exit in RL4, and the collectibles of RL6 and RL10.

Analogously to what was done for the SL mini-game, we also noted the events leading players to retry the level with different settings. The occurrences of an improved, identical or worse performance upon a level retry were counted. Since the level completion is contingent on the exit being found, an improved collectible count was only considered in this analysis when the exit was also reached before and after the repetition. A newly found exit resulted in 9% of retries, while it was lost after 6% of retries. It is expected that in most cases this value would be unchanged since players tended to repeat the level to get more collectibles. For the latter, 17% of the retries resulted in more collectibles and 10% of retries resulted in fewer. It is still noteworthy that 73% of the time the number of collectibles remained unchanged. This suggests that either the settings offered to the player might not be providing enough agency to change this result, or the lack of such affordance is due to the individual design of each level.

5.4. Feedback by Students and Educators

In this section, we analyse the feedback reported by both students and educators to user surveys after these end-users had a chance to play the game in dedicated dissemination events (see Section 4.2). In **Tables 2**, **3**, we report results as the mean and standard deviation from survey results in 5-item Likert scales collected from 130 students and 35 teachers, respectively. The students had not been previously exposed to any AI and ML concepts in their classes. Most of the teachers were teaching language or were primary education teachers. Only 10 of the 35 teachers were specialized in fields such as informatics, robotics, technology, and coding.The teachers and students who responded to the survey had no prior involvement with the game; they had not used the game before or taught concepts related to AI and ML.

It seems that the reception of the game was generally positive by students and educators. Both groups found it relatively easy to understand how to play the game, and reported that they would play it again. Students were slightly more reserved regarding whether they would recommend ArtBot to their friends, while many teachers would recommend it to colleagues for use in the classroom. The responses on how fun the game was were fairly positive from all participants.

The learning aspects of the game were also positively received; all respondents reported that it helped them understand more about AI and ML. Teachers were also very positive regarding whether they believed the game would help students understand the concept of AI. The attitudes toward the implementation of

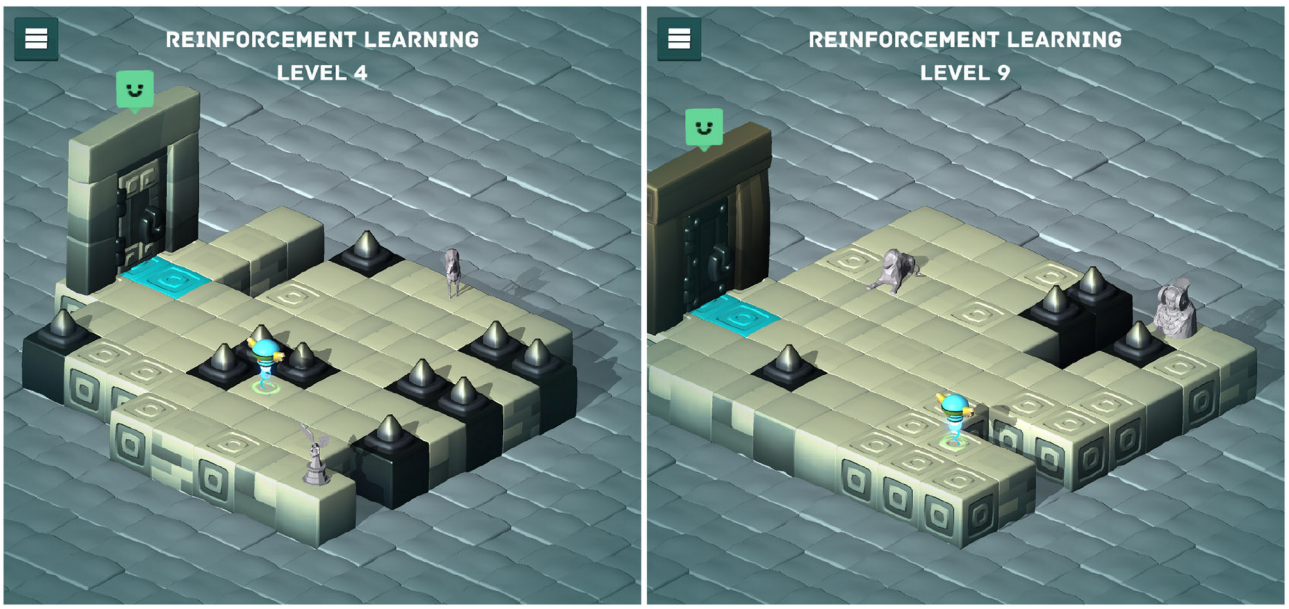


FIGURE 5 | Levels RL4 (left) and RL9 (right) form the Reinforcement Learning mini-game were identified to be the difficult for the players to find the exit and to collect all the treasures, respectively.

TABLE 2 | Likert scale questions asked to the students in the survey. Scores ranged from 1 (strongly disagree) to 5 (strongly agree).

Question to students	Mean	SD	Pos. (%)	Neg. (%)
It was easy to understand how to play the game ArtBot.	4.2	0.8	88	4
It was easy to play the game.	4.2	0.9	79	3
I would play the ArtBot game again.	3.9	1.1	66	9
I thought the ArtBot game was boring.	2.2	1.1	11	64
I would recommend to my friends to play the game ArtBot.	3.5	1.1	58	18
I could learn how to play the game easily.	4.3	0.9	83	5
I thought the ArtBot game was fun.	4.0	1.1	76	9
The ArtBot game helped me understand how Artificial Intelligence works.	4.0	1.0	75	5
In the game, it was easy to understand what machine learning is.	4.0	1.0	75	5
I would like to use the ArtBot game at school.	4.1	1.1	69	9
I would recommend to my teacher to use the ArtBot game in the classroom so as to learn more about Artificial intelligence.	3.7	1.4	68	22

The mean, standard deviation (SD), and the percentage of positive (>3) and negative (<3) replies of the collected values are reported.

the game in the school environment were again positive; students and teachers reported that they would like to use the game in the classroom and, for the teachers, that they intended to use the game in their teaching.

In freeform responses by respondents, one of the most positive aspects cited by students (20 cases) was the learning aspect of the game. Students reported that they enjoyed the combination of learning content and game, they enjoyed the fact that they could learn new concepts through a game, that they had to think, solve problems, and "use their brain". In 17 cases they reported that the game was fun and entertaining, in 12 cases that the graphics and colors were among the positive aspects of the game, and in 4 cases, that they found the game creative. In 15 cases, students reported that they enjoyed their active role, the complexity, and the fact that they had control over the training of the AI agent. Unexpectedly, in 8 cases the learners reported playing with friends as a positive aspect of the game. Since the game is not designed for multiple players, we assume that they referred to the context of playing the game in the same location as their friends; this substantiates the role of the social environment as a motivation for play for young learners (Ferguson and Olson, 2013).

Most of the positive elements of the game reported by teachers were relevant to the learning aspect of ArtBot and its potential to teach students AI and ML concepts and processes through a playful environment (13 cases). The interface, the ease of use, the graphics, and the friendly environment were also among the most cited positive aspects of the game by the teachers (11 cases). Teachers reported that the game environment would attract students and especially students who are already interested in games. Other positive aspects described by the teachers was the explanatory information, the avatar selection, the archaeology aspect, and the fact that the students are manipulating an AI agent.

The most cited negative aspect of the game by students was its pacing; in 9 cases the students described the game as slow, monotonous, time consuming, or boring while in 6 cases, they

TABLE 3 | Likert scale questions asked to the teachers in the survey. Scores ranged from 1 (strongly disagree) to 5 (strongly agree).

Question to teachers	Mean	SD	Pos. (%)	Neg. (%)
It was easy to understand how to play this game.	3.9	0.9	71	3
The game was easy to use.	3.9	0.9	71	3
I would play this game again.	4.0	1.0	74	9
I found the game boring.	2.2	1.2	17	63
I would recommend to my colleagues to use this game in their classroom.	4.1	1.0	77	6
It was easy to learn how to play this game.	3.9	1.0	74	9
I thought the game was fun.	3.8	0.8	74	6
The game helped me understand how Artificial Intelligence works.	4.2	1.0	83	6
It was easy to understand what machine learning is, through this game.	4.2	0.9	86	6
I would like to use this game in the classroom.	4.1	1.0	77	9
I would like to use this game in my teaching in the future.	4.2	0.8	80	3
I believe that this game will help children understand the concept of Artificial intelligence.	4.3	0.7	89	0

The mean, standard deviation (SD), and the percentage of positive (>3) and negative (<3) replies of the collected values are reported.

thought the game was too complex or too difficult to understand and therefore they needed more guidance. In 3 cases the learners reported that they would like to see more levels or more difficult challenges customized to the learners' age. The graphics, the colors, and the avatar were described as negative aspects in 6 cases. The complexity and the difficulty to understand concepts of AI (e.g., the "use experience" parameter) was also highlighted by teachers as a negative aspect (in 7 cases). Teachers reported that more guidelines and tutorials are needed for explaining the process, variables, and concepts to the students. Teachers also listed the quality of the graphics and sound as negative aspects in 3 cases, and in 2 cases teachers suggested that the game should have more variables for the students to manipulate.

The attitudes of teachers and students were generally positive, although there seem to still be some challenges regarding the complexity of the content and the difficulty to address students and teachers of varying levels of AI expertise (see also Zammit et al., 2021).

6. DISCUSSION

The analysis of the data collected has consistently shown a number of important findings regarding the learning and player behavior of the users. The browser platform is evidently very convenient for casual players, as this was by far the most popular for this game. This finding matches our intent to make ArtBot as accessible as possible to a wide audience.

The data indicates that a substantial amount of players meaningfully engaged with both mini-games of ArtBot. Players explored the parameters of the algorithms and were interested to see how manipulating these parameters would vary the outcome of machine learning. The game is therefore successful in its objectives to impart information and awareness about the basics of SL and RL algorithms, their related terminology and processes. However, the lack of interaction with the information screens also shows that it is not evident to users when additional details about the game and background algorithms are made available to them. Based also on feedback by students and teachers in dedicated dissemination events, more effort is needed to better engage the players with the background information and learning content.

The manual image labeling in the SL mini game did not appeal to players, and the outcome of supervised learning does not vary enough with changes in the settings to hold players' attention. In addition, while the first few RL levels were frequently played, the subsequent ones did not offer enough variety or novelty to retain player interest. This could hinder our goal to disseminate the game to a wide audience, and to enhance the learning process by active participation through game-based learning. To address this, additional prompts or datasets closer to the learners' interests could better indicate the role and impact of labeling on the training of the model and the behavior of the agent in the SL process. Similarly, the RL levels could be fine-tuned to provide settings and agent behaviors which are varied and obvious to the players. Additionally, interaction data have brought to light specific design issues with individual levels, such as RL4 and RL9, which require further tuning to align their difficulty to the intended difficulty progression in the level order.

The game-based learning aspects of ArtBot in formal education was positively received by both students and teachers. Both end-users mostly agreed that the implementation of ArtBot in the classroom for teaching and learning about AI and ML would motivate and engage the students, and could play a more active role in their learning. We note that the sample may be biased, especially for the educators, since the survey was completed by teachers who chose to participate in the relevant events; these teachers had, most probably, positive attitudes toward new tools and new concepts. That said, we can assume that games—and particularly ArtBot—can be a useful tool for educators planning to introduce AI education and literacy in their classroom.

6.1. Limitations

In this paper, we explored how anonymous usage data and surveys targeting educators and students can be used to gauge how a game designed for imparting AI and ML concepts was perceived by a broad audience. As such, our findings are specific to the game ArtBot and certainly not generalisable to all games on AI and ML. However, this paper aims to provide insights on the design and analytics of games aiming to teach AI and ML concepts to young students, and to highlight the potential of games to teach AI and ML.

On that note, ArtBot is intended to be used by teachers and students in a classroom. Since the data collection process is anonymous, no information about the players' ages or their roles

is available. This gives limited context in which to evaluate the interactions with the game. For example, it would have been interesting to understand which age groups were playing the later levels, and which types of end-users (teachers vs. students) were accessing the additional information.

During the design process of ArtBot, we strove to appeal to a broad age group by using simple graphics, familiar controls, and gameplay that has an immediate reward but can be explored deeper in accordance to the curiosity and understanding of the player. Our objective was a game that is understandable enough for primary school students, yet one that can still offer a challenge to secondary level ones. This versatility comes at the cost that ArtBot can not fully address the distinct pedagogical requirements of one specific group over the other. The game therefore trades off a more targeted teaching approach for a wider reach. This was corroborated by responses from educators in our surveys, as teachers of different topics and at different educational institutions and levels gave very different directions toward improvement of the game.

The browser platform, despite the advantages of neither requiring any installation nor a specific device, posed significant problems in the analysis. Users and sessions cannot be clearly distinguished, especially if users close or reopen their browsers, use different browsers, or even multiple tabs. Furthermore, it is very easy for users to switch to different tabs and leave the game running in the background, returning to it and continuing later. This behavior hinders the analysis due to the misleading timestamps corresponding to the same user and session. Moreover, the number of unique users is likely inflated due to the above behavior and/or the use of incognito windows and cookie blockers.

Regarding the attainment of the learning objectives and understanding of the AI and ML concepts addressed in the game, both teachers and students reported a positive impact in dedicated feedback sessions. However, further tests that combine quantitative (e.g., pre- and post-tests) and qualitative (e.g., interviews) analysis are needed to more objectively examine the learning impact of the game and identifying potential misconceptions.

6.2. Future Work

The analysis reported in Section 5 clearly outlined areas where the game can be improved. The SL mini-game can be reformatted to give more weight to user settings, by perhaps introducing different parameters of the algorithm which have a more drastic effect on the learning process, and thereby relying less on image labeling. The RL levels require additional features and a gradual increase in difficulty, with perhaps a better reward system for players, such as a points system with a leader board showing best results, or a list of achievements for the player to accomplish. The additional information button can be highlighted whenever it is populated with new information, or it could be shown automatically the first time that new information is available to the players, in the fashion of a game tutorial. In the latter case, the text would be revised and made less verbose.

Externally to the game, teaching resources and accompanying materials have been developed as classroom aids to enhance the learning experience with ArtBot[11]. This material can be further developed to address some of the shortcomings identified, namely additional information about the underlying algorithms and their use in different real-world applications. The feedback of both teachers and students reported in Section 5.4 will be an important guide toward improving the educational material to mitigate some of the difficulties in understanding the underlying algorithms and to better connect it to everyday ML uses in students' lives.

7. CONCLUSION

ArtBot is part of the toolkit developed through the LearnML project to support experimentation, reflection, and critical thinking about AI and ML to primary and secondary school students. Its goal was to teach the basics of supervised learning and reinforcement learning through a playful and exploratory experience. This paper analyzed how ArtBot has been used by different players since its launch in April of 2021, including the feedback of teachers and students in dedicated playtesting events. Through anonymously collected usage metrics, we identify the usability of the game, its user interface, and design effectiveness. The data revealed that the game was generally well received, having over 2,000 unique users, with the browser version being the most popular platform. This indicates that our efforts to make the game easily accessible were fruitful. The players were largely successful completing the in-game activities; many players explored the various ML parameter setups, but only a few explored additional information to learn more about the topics of ML. The dedicated feedback by students and teachers also indicated a generally positive outlook on the use of ArtBot in the classroom, but also raised concerns regarding the game's pacing and the complexity of some of the concepts introduced. A number of potential areas of improvement were identified, both in broad scope as well as specific design tweaks for each portion of the game. With these findings in hand, the game can be refined to enhance player engagement, and to maximize the benefits of a game-based learning experience in the classroom.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data has been collected anonymously but pertains to the users who have played the game. Although specific users cannot be identified from the data, we do not wish for the data to be publicly accessible. Requests to access the datasets should be directed to MZ, marvin.zammit@um.edu.mt.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next

[11] The LearnML Guidebook is freely available at http://art-bot.net/teachers/.

of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MZ and IV carried out the data collection and analysis reported in the paper. MZ, IV, and AL each contributed to the writing in the various sections of the paper. AL and GY advised on the research direction and the text, and oversaw the implementation, analysis, and authoring process. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Learn to Machine Learn (LearnML) project, under the Erasmus+ Strategic Partnership program (Project Number: 2019-1-MT01-KA201-051220).

ACKNOWLEDGMENTS

The authors would like to thank the many anonymous users of ArtBot, as well as all partners of the LearnML project for their feedback throughout the project and their participation in dissemination events for promoting ArtBot to the general public.

REFERENCES

Clark, D., Nelson, B., Sengupta, P., and D'Angelo, C. (2009). "Rethinking science learning through digital games and simulations: genres, examples, and evidence," in *Proceedings of the Workshop on Learning Science: Computer Games, simulations, and Education* (Washington, DC).

Clarke, D., and Noriega, L. (2003). "Games design for the teaching of artificial intelligence," in *Interactive Convergence: Research in Multimedia* (Prague).

Druga, S., Williams, R., Breazeal, C., and Resnick, M. (2017). ""Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction," in *Proceedings of the Conference on Interaction Design and Children* (New York, NY), 595–600. doi: 10.1145/3078072.3084330

Ferguson, C. J., and Olson, C. K. (2013). Friends, fun, frustration and fantasy: Child motivations for video game play. *Motivat. Emot*. 37, 154–164. doi: 10.1007/s11031-012-9284-7

Giannakos, M., Voulgari, I., Papavlasopoulou, S., Papamitsiou, Z., and Yannakakis, G. (2020). "Games for artificial intelligence and machine learning education: review and perspectives," in *Non-Formal and Informal Science Learning in the ICT Era*, ed M. Giannakos (Cham: Springer), 117–133. doi: 10.1007/978-981-15-6747-6_7

Glaser, M., Garsoffky, B., and Schwan, S. (2009). Narrative-based learning: possible benefits and problems. *Communications* 34, 429–447. doi: 10.1515/COMM.2009.026

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., and Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Comput. Educ*. 102, 202–223. doi: 10.1016/j.compedu.2016.09.001

Hartness, K. (2004). Robocode: using games to teach artificial intelligence. *J. Comput. Sci. Coll*. 19, 287–291. doi: 10.5555/1050231.1050275

Kafai, Y. B., and Burke, Q. (2015). Constructionist gaming: understanding the benefits of making games for learning. *Educ. Psychol*. 50, 313–334. doi: 10.1080/00461520.2015.1124022

Lai, E. R. (2011). Critical thinking: a literature review. *Pearsons Res. Rep*. 6, 40–41.

Mertala, P. (2019). Young children's conceptions of computers, code, and the Internet. *Int. J. Child Comput. Interact*. 19, 56–66. doi: 10.1016/j.ijcci.2018.11.003

Parker, J., and Becker, K. (2014). "ViPER: Game that teaches machine learning concepts - a postmortem," in *Proceedings of the IEEE Games and Entertainment Media Conference* (Toronto, ON).

Quinlan, J. (1992). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine behaviour. *Nature* 568, 477–486. doi: 10.1038/s41586-019-1138-y

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA; London: MIT Press.

Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., Valtonen, T., and Vartiainen, H. (2020). "Co-designing machine learning apps in K-12 with primary school children," in *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (Los Alamitos, CA), 308–310. doi: 10.1109/ICALT49669.2020.00099

Tsai, Y.-L., and Tsai, C.-C. (2020). A meta-analysis of research on digital game-based science learning. *J. Comput. Assist. Learn*. 36, 280–294. doi: 10.1111/jcal.12430

Turchi, T., Fogli, D., and Malizia, A. (2019). Fostering computational thinking through collaborative game-based learning. *Multimedia Tools Appl*. 78, 13649–13673. doi: 10.1007/s11042-019-7229-9

Vartiainen, H., Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., and Valtonen, T. (2021). Machine learning for middle schoolers: learning through data-driven design. *Int. J. Child Comput. Interact*. 29. doi: 10.1016/j.ijcci.2021.100281

Voulgari, I. (2020). "Digital games for science learning and scientific literacy," in *Non-Formal and Informal Science Learning in the ICT Era*, ed M. Giannakos (Cham: Springer Nature), 35–49. doi: 10.1007/978-981-15-6747-6_3

Voulgari, I., Zammit, M., Stouraitis, E., Liapis, A., and Yannakakis, G. N. (2021). "Learn to machine learn: designing a game based approach for teaching machine learning to primary and secondary education students," in *Proceedings of the ACM Interaction Design and Children Conference* (New York, NY: Association for Computing Machinery), 593–598. doi: 10.1145/3459990.3465176

Watkins, C. J. C. H., and Dayan, P. (1992). Q-learning. *Mach. Learn*. 8, 279–292. doi: 10.1023/A:1022676722315

Yannakakis, G. N., and Togelius, J. (2018). *Artificial Intelligence and Games*. (Cham: Springer). doi: 10.1007/978-3-319-63519-4

Zammit, M., Voulgari, I., Liapis, A., and Yannakakis, G. N. (2021). "The road to AI literacy education: from pedagogical needs to tangible game design," in *Proceedings of the European Conference on Games Based Learning* (Reading: Academic Conferences International Limited), 763–771.

Zhou, Z., and Liu, S. (2021). *Machine Learning*. Cham: Springer. doi: 10.1007/978-981-15-1967-3

Zimmermann-Niefield, A., Turner, M., Murphy, B., Kane, S. K., and Shapiro, R. B. (2019). "Youth learning machine learning through building models of athletic moves," in *Proceedings of the ACM International Conference on Interaction Design and Children* (Reading: Academic Conferences International Limited), 121–132. doi: 10.1145/3311927.3323139

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Psychology

METHODS
published: 01 August 2022
doi: 10.3389/fpsyg.2022.846621

A Knowledge Query Network Model Based on Rasch Model Embedding for Personalized Online Learning

Yan Cheng[1,2], Gang Wu[1], Haifeng Zou[1], Pin Luo[1] and Zhuang Cai[1]*

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [2] Jiangxi Provincial Key Laboratory of Intelligent Education, Nanchang, China

The vigorous development of online education has produced massive amounts of education data. How to mine and analyze education big data has become an urgent problem in the field of education and big data knowledge engineering. As for the dynamic learning data, knowledge tracing aims to track learners' knowledge status over time by analyzing the learners' exercise data, so as to predict their performance in the next time step. Deep learning knowledge tracking performs well, but they mainly model the knowledge components while ignoring the personalized information of questions and learners, and provide limited interpretability in the interaction between learners' knowledge status and questions. A context-aware attentive knowledge query network (CAKQN) model is proposed in this paper, which combines flexible neural network models with interpretable model components inspired by psychometric theory. We use the Rasch model to regularize the embedding of questions and learners' interaction tuples, and obtain personalized representations from them. In addition, the long-term short-term memory network and monotonic attention mechanism are used to mine the contextual information of learner interaction sequences and question sequences. It can not only retain the ability to model sequences, but also use the monotonic attention mechanism with exponential decay term to extract the hidden forgetting behavior and other characteristics of learners in the learning process. Finally, the vector dot product is used to simulate the interaction between the learners' knowledge state and questions to improve the interpretability. A series of experimental results on 4 real-world online learning datasets show that CAKQN has the best performance, and its AUC value is improved by an average of 2.945% compared with the existing optimal model. Furthermore, the CAKQN proposed in this paper can not only track learners' knowledge status like other models but also model learners' forgetting behavior. In the future, our research will have high application value in the realization of personalized learning strategies, teaching interventions, and resource recommendations for intelligent online education platforms.

Keywords: personalized education, deep learning, knowledge tracking, forgetting behavior, interpretability

OPEN ACCESS

Edited by:
Manuel Gentile, Istituto per le Tecnologie Didattiche (ITD) (CNR), Italy

Reviewed by:
Sílvio Manuel da Rocha Brito, Instituto Politécnico de Tomar (IPT), Portugal
David Paulo Ramalheira Catela, Polytechnic Institute of Santarém, Portugal

****Correspondence:***
Yan Cheng
chyan88888@jxnu.edu.cn

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *31 December 2021*
Accepted: *23 May 2022*
Published: *01 August 2022*

Citation:
Cheng Y, Wu G, Zou H, Luo P and Cai Z (2022) A Knowledge Query Network Model Based on Rasch Model Embedding for Personalized Online Learning. Front. Psychol. 13:846621. doi: 10.3389/fpsyg.2022.846621

INTRODUCTION

With the rapid development of Internet technology and artificial intelligence technology in the field of education, online learning platforms such as massive open online courses (MOOCs) have become increasingly popular. Learners' activities on online learning platforms have generated massive amounts of educational data. How to mine and analyze large amounts of educational data has become an urgent problem in the field of education and big data knowledge engineering (Hu et al., 2020). Since learners' behavior, knowledge state, and psychological factors in the learning process are the key factors for evaluating their learning effectiveness (Yang and Li, 2018), and these factors are constantly changing over time, it is of great significance to construct a learner model oriented to dynamic learning data.

Different from the cognitive diagnosis model (CDM) for static learning data, knowledge tracing (KT) aims to dynamically track learners' knowledge status over time by analyzing the learners' historical exercise data, so as to predict their performance in the next time step. The learner's historical exercise data is a sequence composed of the questions, the knowledge components (KCs) contained in the questions, and the learner's answers (Liu et al., 2021). The three core elements of questions, KCs and learners constitute the three basic objects of the KT data processing, the interaction between them is shown in **Figure 1**. KT is the quantitative analysis and modeling of the relationship between three types of objects. For example, the prediction of students' knowledge mastery state is to calculate the mastery probability between "students and knowledge" by using the interaction between "students and problems" and the correlation information between "problems and knowledge." The interaction between different objects is the main information used in the KT modeling process (Sun et al., 2021). Therefore, the KT model not only needs to accurately assess the learner's knowledge state and predict their answer in the future but also needs to provide explanations for the interaction between different objects (Hu et al., 2020).

Traditional KT methods mainly include Bayesian knowledge tracking (BKT) (Corbett and Anderson, 1994) based on hidden Markov model (HMM) (Rabiner and Juang, 1986) and item response theory (IRT) (Fan, 1998). In recent years, researchers have tended to use more complex and flexible models like deep networks to make full use of hidden information in large-scale learner response datasets. The deep knowledge tracing (DKT) (Piech et al., 2015) model introduced recurrent neural network (RNN) into the KT field for the first time and achieved success. Compared with the traditional KT model, the predictive ability based on the deep learning method has been significantly improved. However, most of the current KT methods based on deep learning mostly use KCs to index questions, ignoring the rich information contained in the questions and the context. For example, investigating different questions of the same KC may cause individual differences between questions due to different difficulty settings. In addition, the personalized interaction between learners' knowledge status and questions representation is often overlooked, which leads to poor interpretability of the KT method based on deep learning. In response to the above problems, we propose a context-aware attentive knowledge query network (CAKQN) model based on the embedded Rasch model, which is the single parameter IRT model. First, input the learner interaction tuple and questions into the embedded component based on the Rasch model to obtain personalized representations of the learner interaction tuple and questions, and capture the characteristics of individual differences between different questions containing the same KC and the learners' personal abilities. Next, based on the definition of memory trace decline in educational psychology theory (Bailey, 1989) that human memory fades automatically over time, a network structure of long short-term memory network + monotonic attention mechanism is designed to learn personalized learner knowledge state and context-aware representation of the questions. The learning process of learners is continuous, so the sequence structure of learning records cannot be destroyed in the KT modeling process. The structure we designed uses a monotonic attention mechanism with an exponential decay term to reduce the importance of learner interaction tuples in the distant past without destroying the sequence structure of the learners' historical learning records, and it can extract features such as forgetting behavior that exist in the learning process of learners. Finally, based on the fact that learners answer questions based on their knowledge status and personal abilities, the vector dot product is used to simulate the personalized interaction between learners' knowledge status and questions to improve the interpretability of the model. We used four publicly available real online education datasets to evaluate the model. Experiments show that the CAKQN model has the best performance, and its AUC value is 2.945% higher than the existing optimal model on average. In addition, our paper also conducted a series of ablation analysis and knowledge tracking visualization experiments to verify the excellent interpretability and personalization capabilities of the CAKQN model. In the future, our research will have high application value in the realization of personalized learning strategies, teaching interventions, and resource recommendations for intelligent online education platforms.

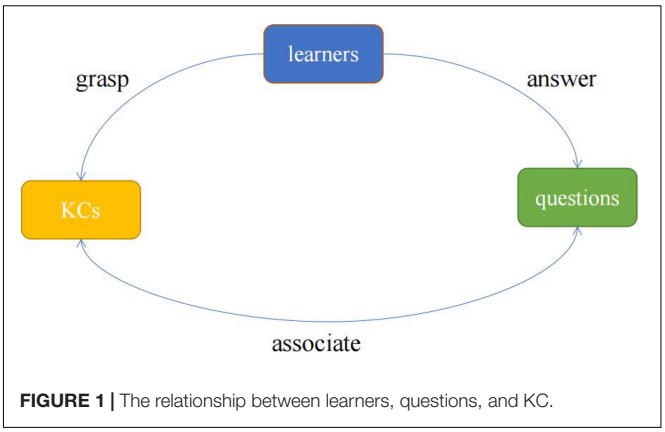


FIGURE 1 | The relationship between learners, questions, and KC.

RELATED WORK

Traditional Knowledge Tracking Methods

Traditional knowledge tracking methods are mainly divided into two categories: IRT and BKT, and IRT is one of the important psychological and educational theories (Cheng et al., 2019). The single-parameter IRT model (i.e., Rasch model) outputs the probability of learners answering the items correctly during the test according to the learner's ability level and the difficulty level of the items (i.e., questions). The probability is defined by the item response function with the following characteristics: if the learner's ability level is higher, the learner has a higher probability of answering an item correctly. Conversely, if an item is more difficult, the probability of the learner answering the item correctly is lower. The item response function is defined as follows:

$$P(a) = \sigma\left(\theta - \beta_j\right) = \frac{1}{1 + e^{-D(\theta - \beta_j)}} \tag{1}$$

The more complex two-parameter item response function introduces item discrimination α_j, which is defined as follows:

$$P(a) = \sigma\left(\theta - \beta_j\right) = \frac{1}{1 + e^{-D\alpha_j(\theta - \beta_j)}} \tag{2}$$

Where σ is the sigmoid function, D is a constant, usually set to 1.7, θ is learner's ability level, β_j is the difficulty level of item j. Since the IRT model was originally designed for educational testing environments, the model assumes that learners' abilities remain unchanged during the testing process. In reality, the knowledge state of learners changes with time step, so it cannot be directly applied to KT tasks.

The BKT model updates the learner's knowledge state through HMM modeling, and predicts the learner's performance at the next time step accordingly. However, many simplified assumptions used in the BKT model are impractical. One of them is that all learners and questions containing the same KC are considered the same. Therefore, the researchers studied various personalizations of the BKT model. Some researchers endow the BKT model with personalized capabilities on specific parameters of KC (Pardos and Heffernan, 2011) and specific parameters of learners (Yudelson et al., 2013). Some other researchers have also studied the synthesis of the BKT model and the IRT model (Khajah et al., 2014; Wilson et al., 2016) to enhance the model's personalization ability when dealing with questions and learners. However, such expansion usually requires a lot of feature engineering work and will result in a significant increase in computing requirements.

Deep Learning Knowledge Tracking

In recent years, deep learning has attracted attention from researchers with its powerful feature extraction capabilities. Many researchers have applied it to the KT field, which is called DLKT (deep learning knowledge tracing) (Liu et al., 2021). Compared with BKT and IRT, DLKT does not require manually annotated KC information and can capture more complex learner knowledge representations from large-scale learner response datasets. DKT and dynamic key-value memory network (DKVMN) (Zhang et al., 2017) have shown strong predictive ability in predicting learners' future performance, and have become the benchmark for subsequent DLKT methods. DKT takes the learner's historical learning interaction sequence as input, then uses RNN to encode it into the learner's knowledge state, and finally inputs it into a linear layer activated by a Sigmoid function to get the prediction result. DKT, which simply represents the learner's knowledge state as a vector, while DKVMN uses a static external matrix to store KC and uses a dynamic matrix to update the learner's mastery of KC. However, the simple splicing between the two vectors representing the learner's knowledge state and KC in the DKVMN model is not enough to explain the process of interaction between the learner's knowledge state and the KC contained in the question (Daniluk et al., 2017). The knowledge query network (KQN) (Lee and Yeung, 2019) model uses the vector dot product to more accurately simulate the interaction between the learner's knowledge state and KC, and achieves better results. Self-attentive knowledge tracing (SAKT) (Pandey and Karypis, 2019) model is the first to use the Transformer structure in the KT field to replace RNN to automatically focus on the record of questions in the learner's historical interaction sequence that has a greater impact on the prediction results and achieves model performance. The substantial increase. However, the above models use KCs to index questions, that is, all different questions containing the same KC are regarded as equivalent. This way ignores the rich information contained in the question itself and the context. Context-aware attention knowledge tracing (AKT) (Ghosh et al., 2020). The framework based on the SAKT model uses the Rasch model to regularize concept and question embeddings. These embeddings can capture questions that contain the same KC, without using too many parameters. In addition, AKT also uses a new monotonic attention mechanism to link learners' future responses to questions with their historical interaction sequences to extract features such as hidden forgetting behavior in the learning process of learners. However, the AKT model also uses unreasonable vector simple splicing to simulate learner knowledge status and question interaction, and it loses the ability to model sequence due to the Transformer structure like SAKT.

Considering the advantages and disadvantages of KQN model and AKT model, this paper proposes a context-aware knowledge query network (CAKQN) based on Rasch model embedding. It not only retains the ability of model sequence but also obtains personalized contextual representations of questions and learners. We improve the model's performance in predicting future learner responses. Moreover, the interpretability of the model in terms of learner knowledge status and questions interaction is enhanced.

OUR PROPOSED METHOD

This section first introduces the problem setup of knowledge tracing and the symbolic representation of related concepts, then introduces the difference between ordinary attention mechanism and monotonic attention mechanism with exponential decay, and then describes the overall context-aware knowledge query

network model based on Rasch model embedding framework, and finally introduce each component of the model and its loss function in turn.

Knowledge Tracing Problem Setup

Assuming that there are M questions and N KCs in the original dataset, each learner's interaction record is composed of the learner's long questions and responses at each time step. For the learner i at time step t, a learner interaction tuple $x_t = (q_t^i, c_t^i, r_t^i)$ is composed of: the question q_t^i he or she answered, the KC c_t^i covered by the question, and the learner's response r_t^i to the question. Where q_t^i is the question index, $q_t^i \in \{1, \cdots, M\}$, c_t^i is the KC index, $c_t^i \in \{1, \cdots, N\}$, and r_t^i is the response, $r_t^i = \{0, 1\}$. Under this notation, $(q_t, c_t, 1)$ means learner i responded to question q_t on concept c_t correctly at time step t. This setting is different from some previous deep knowledge tracking work, which often ignores the question index and set the learner's interaction tuple as (c_t^i, r_t^i). For convenience, the superscript i is omitted in the following discussion. Therefore, given learner's historical learning interaction sequence $X_t = \{x_1, x_2, \cdots, x_t\}$ at time step t and question q_{t+1} on concept c_{t+1} at time step $t+1$, the goal of the KT model is to find the probability $P(r_{t+1} = 1|X_t, q_{t+1}, c_{t+1})$.

Monotonic Attention Mechanism With Exponential Decay

Under the ordinary dot product attention mechanism, the input is mapped to three vectors: *Query*, *Key*, and *Value* by embedding layer, and values of dimension $D_q = D_k$, D_k and D_v. Let $q_t \in \mathbb{R}^{D_k \times 1}$ donate the *Query* corresponding at time step t, the calculation formula of the scaled dot product attention value $\alpha_{t,\tau}$ normalized by the softmax function is:

$$\alpha_{t,\tau} = Softmax(\frac{q_t^T k_\tau}{\sqrt{D_k}}) = \frac{exp(\frac{q_t^T k_\tau}{\sqrt{D_k}})}{\sum_{\tau'} exp(\frac{q_t^T k_\tau}{\sqrt{D_k}})} \in [0, 1] \quad (3)$$

Where $k_\tau \in \mathbb{R}^{D_k \times 1}$ donate *Key* at time step τ .

However, this ordinary zoom dot product attention mechanism is not enough for KT tasks. The reason is that learners have forgetting behaviors in the learning process, and learners will have memory decline in the real world (Pashler et al., 2009). In other words, when the model predicts the learner's reaction to the next question, his performance in the distant past is not as important as his recent performance. Therefore, Ghosh et al. (2020) add a multiplicative exponential decay term to the attention scores. So the calculation of the new monotonic attention mechanism is as follows:

$$\alpha'_{t,\tau} = \frac{exp(s_{t,\tau})}{\sum_{\tau'} exp(s_{t,\tau'})} \quad (4)$$

$$s_{t,\tau} = \frac{exp(-\theta \cdot d(t, \tau)) \cdot q_t^T k_\tau}{\sqrt{D_k}} \quad (5)$$

Where $\theta > 0$ is a learnable decay rate parameter, and $d(t, \tau)$ is temporal distance measure between time steps t and τ. In other words, the attention weight of the current question to the past question not only depends on the similarity between the corresponding sums, but also depends on the relative time steps between them. The calculation method of $d(t, \tau)$ is as follows:

$$d(t, \tau) = |t - \tau| \cdot \sum_{t'=\tau+1}^{t} \gamma(t, t') \quad (6)$$

$$\gamma(t, t') = \frac{exp(\frac{q_t^T k_{t'}}{\sqrt{D_k}})}{\sum_{1 \le \tau' \le t} exp(\frac{q_t^T k_{\tau'}}{\sqrt{D_k}})} \quad (7)$$

The calculation formula of the final output of the monotonic attention mechanism is as follows:

$$Monotonic\_Attention(Query, Key, Value) = \sum_{\tau=1}^{t} \alpha'_{t,\tau} v_\tau \quad (8)$$

Where $v_\tau \in \mathbb{R}^{D_k \times 1}$ donate *Key* at time step τ .

Model Framework

This paper proposes a context-aware knowledge query network based on Rasch model embedding. **Figure 2** shows the overall framework of the model. It contains 4 components: *Embedded Layer Based on Rasch Model*, *Knowledge State Encoder*, *Question Encoder*, and *Knowledge Status Query*.

(1) *Embedded Layer Based on Rasch Model*: Get the personalized embedding of the learner interaction tuple at the current time step and the next-time step question, and capture the characteristics of individual differences between different questions on the same KC and the learners' personal abilities.

(2) *Knowledge State Encoder*: First, use the location information provided by the long short-term memory network to model the context of the learner's historical interaction sequence, and retain the ability of the model to model the sequence. Then, the monotonic attention mechanism with exponential decay term is used to reduce the importance of learner interaction tuples in the distant past, extract the forgetting behavior and other characteristics of learners in the learning process, and obtain the contextual perception vector of the learner's knowledge state at the current time step.

(3) *Question Encoder*: It is exactly the same as the network structure adopted by the knowledge state encoder to obtain the context awareness vector of the question at the current time step.

(4) *Knowledge Status Query*: The dot product operation is performed on the vector representing the learner's knowledge state and the question at the current time step to simulate the interaction between the learner's knowledge state and the question, and the result of the dot product is input into the sigmoid function to obtain the final prediction of the probability that the learner will answer correctly at the next time step.

Embedding Layer Based on Rasch Model

Existing KT methods mostly use KC to index questions, that is, set $q_t = c_t$, because the number of questions in the real world is far greater than the number of KC, so using KC to index

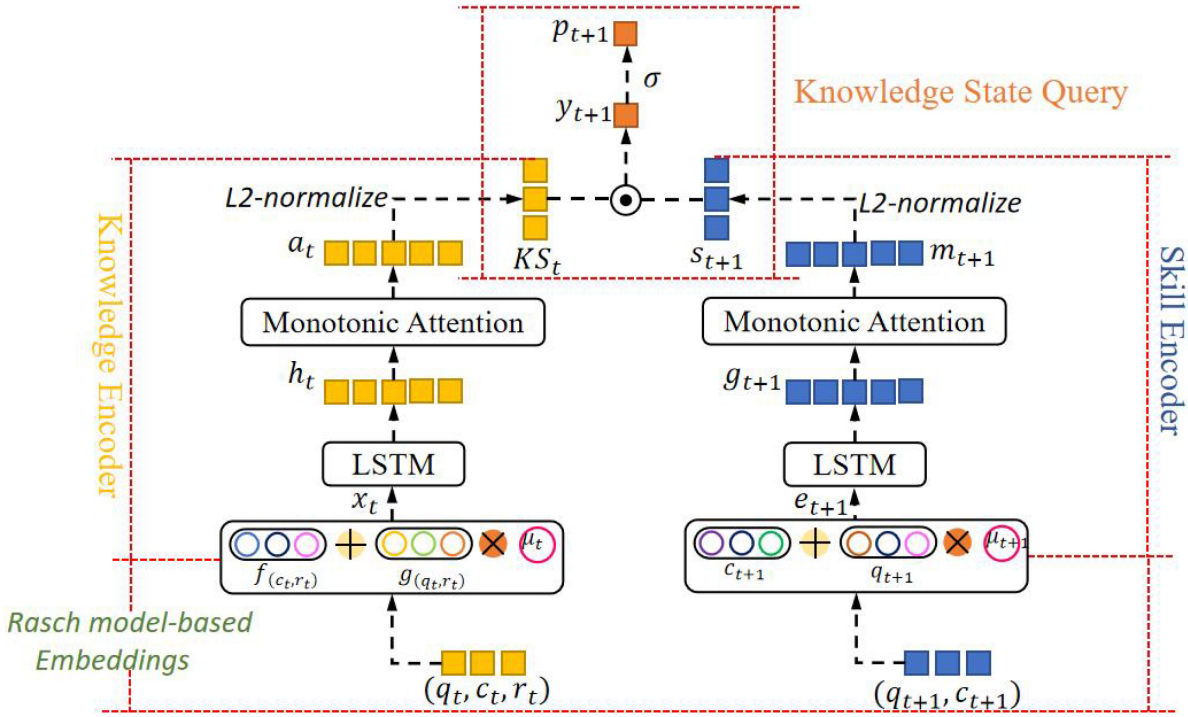


FIGURE 2 | The overall framework of the CAKQN model.

questions can effectively avoid over-parameterization and over-fitting. However, this setting ignores the individual differences between question covering the same KC, and limits the flexibility of the KT method and its ability to be personalized.

This article uses the classic Rasch model in psychometric theory to construct learner interaction tuples and question embedding. There are two important parameters in the Rasch model: the difficulty of the question and the ability of the learners. Therefore, at time step t, the final embedded representation of the learner's interaction tuple is expanded to:

$$x_t = f_{(c_t,r_t)} + \mu_t \cdot g_{(q_t,r_t)} \tag{9}$$

Where $f_{(c_t,r_t)} \in \mathbb{R}^D$, $g_{(q_t,r_t)} \in \mathbb{R}^D$, they respectively, represent the embedding vector of the KC response tuple and the embedding vector of the question response tuple. And μ_t is a learnable scalar, which represents the learner's ability parameter. At the next time step, the final embedded representation of the question is expanded to:

$$e_{t+1} = c_{t+1} + \mu_{t+1} \cdot q_{t+1} \tag{10}$$

Where $c_{t+1} \in \mathbb{R}^D$ is the embedding vector of KC contained in this question, $q_{t+1} \in \mathbb{R}^D$ is the embedding vector of the question. And μ_{t+1} is also a learnable scalar, it represents the difficulty parameter, which controls the degree of deviation of the question from the KC contained in it. These Rasch model-based embeddings strike an appropriate balance between obtaining personalized representations and avoiding excessive parameterization.

Knowledge State Encoder

In the *Knowledge State Encoder*, the structure of the LSTM layer + monotonic attention mechanism layer is used to obtain the context perception results of learner interaction sequences. The way learners understand and learn when answering questions is based on their own knowledge state, and the learner's knowledge state is related to the learner's historical learning interaction sequence. For two learners with different historical learning interaction sequences, the way they understand the same question and the knowledge they gain from the exercise may be different. Therefore, we use the LSTM structure to ensure that the original learner history learning interaction sequence is not destroyed on the time scale, and introduce the monotonic attention mechanism to summarize the performance of the past learners in the correct time range, tap the hidden features of the learning process, and then obtain their knowledge state. Given input x_t, the knowledge state encoder first inputs it to the LSTM layer to obtain its hidden state h_t. Then input h_t to the monotonic attention mechanism layer to get the weighted vector a_t, and finally a through a fully connected layer and L2 normalization to get the final output knowledge state vector KS_t. The calculation process is as follows:

$$\begin{cases} i_t = \sigma\left(W_i\ \left[x_t,\ h_{t-1},\ c_{t-1}\right] + b_i\right) \\ f_t = \sigma(W_f[x_t,\ h_{t-1},\ c_{t-1}] + b_f) \\ o_t = \sigma(W_o[x_t,\ h_{t-1},\ c_{t-1}] + b_o) \\ c_t = f_t c_{t-1} + i_t \tanh\left(W_c\ \left[x_t,\ h_{t-1}\right] + b_c\right) \\ h_t = o_t tanh(c_t) \end{cases} \tag{11}$$

Where i_t, f_t, o_t, c_t are the input gate, forget gate, output gate and unit state, respectively.

$$a_t = Monotonic\_Attention(x_t, x_t, h_t) \tag{12}$$

$$KS_t = L2\_normalize(W_{h,KS}a_t + b_{h,KS}) \tag{13}$$

Where $W_{h,KS} \in \mathbb{R}^{d \times H_{LSTM}}$, $b_{h,KS} \in \mathbb{R}^d$, and H_{LSTM} is the size of the hidden layer of the LSTM, d is the dimension of the knowledge state vector KS_t and the question vector S_{t+1}. *L2_normalize*is L2 normalization, the reason for this limitation is to allow the knowledge state vector and the question vector to be a dot product. In addition, in order to avoid overfitting, regularization is used in the output layer of LSTM.

Question Encoder

In this article, the question encoder uses the same network structure as the knowledge state encoder, and the purpose is also to capture the context-aware results of the question at the next time step. The specific calculation process of the input question embedding e_{t+1} to obtain the question vector s_{t+1} by the question encoder is as follows:

$$\begin{cases} i_t = \sigma\left(W_i\ \left[e_{t+1},\ g_{t-1},\ c_{t-1}\right] + b_i\right) \\ f_t = \sigma(W_f[e_{t+1},\ g_{t-1},\ c_{t-1}] + b_f) \\ o_t = \sigma(W_o[e_{t+1},\ g_{t-1},\ c_{t-1}] + b_o) \\ c_t = f_t c_{t-1} + i_t \tanh\left(W_c\ \left[e_{t+1},\ g_{t-1}\right] + b_c\right) \\ g_{t+1} = o_t tanh(c_t) \end{cases} \tag{14}$$

$$m_{t+1} = Monotonic\_Attention(e_{t+1}, e_{t+1}, g_{t+1}) \tag{15}$$

$$s_{t+1} = L2\_normalize(W_{h,KS}m_{t+1} + b_{h,KS}) \tag{16}$$

Knowledge Status Query

Do the dot product operation on the dimensional knowledge state vector KS_t and the dimensional question vector S_{t+1} obtained by the knowledge state encoder and the item encoder, respectively, and then input the result into the sigmoid activation function to obtain the final prediction of the probability p_{t+1} that the learner answers the next question correctly. Calculated as follows:

$$y_{t+1} = KS_t \cdot S_{t+1} \tag{17}$$

$$p_{t+1} = \sigma(y_{t+1}) \tag{18}$$

The dot product of knowledge state vector and question vector conforms to the process of real world middle school learners answering questions based on their own knowledge state (Lee and Yeung, 2019), which makes the model more explanatory.

Optimization

We use the backpropagation algorithm to train the network model, and update the model parameters by minimizing the cross entropy loss of the prediction probability and the labeled result. At each time step t, calculate the cross entropy loss result of a single learner, and sum the $t = 1, \cdots, T-1$ loss of all learners to get the total loss. The specific calculation process is:

$$\ell(\theta_{model}|r^i_{t+1}, p^i_{t+1}) = -[r^i_{t+1} log p^i_{t+1} + (1 - r^i_{t+1}) log(1 - p^i_{t+1})] \tag{19}$$

$$\mathcal{L}(\theta_{model}|r_{2:t+1}, p_{2:t+1}) = \sum_i \sum_{t=1}^{T-1} \ell(\theta_{model}|r^i_{t+1}, p^i_{t+1}) \tag{20}$$

TABLE 1 | Statistics of dataset.

Dataset	learners	KCs	Questions	Responses
ASSISTments2009	4,151	110	16,891	325,637
ASSISTments2015	19,840	100	–	683,801
ASSISTments2017	1,709	102	3,162	942,816
Statics2011	333	1,223	–	189,297

EXPERIMENTS

In this section, we first introduce the details of the dataset, experimental parameter settings and evaluation indicators, and then show the performance of this model and other models in 4 real-world online education datasets. Finally, we use ablation experiments to further verify the effectiveness of the Rasch model-based embedding, monotonic attention mechanism and question context-aware representation.

Datasets

We used four publicly available real online education datasets to evaluate the model, namely ASSISTments2009, ASSISTments2015, ASSISTments2017[1], and Statics2011[2]. The ASSISTments datasets are collected from the ASSISTments online tutoring platform. And the ASSISTments2009 dataset has been the accepted standard dataset of the KT method for the past 10 years. The Statics2011 dataset was collected from a university-level statics engineering course. In all datasets, the preprocessing steps in this paper follow a series of standards in Ghosh et al. (2020). In **Table 1**, we list the number of learners, KCs (i.e., concepts, knowledge points), questions, and learner interaction tuples. In these datasets, only the ASSISTments2009 and ASSISTments2017 datasets contain question IDs. Therefore, the model based on the Rasch model embedding is only applicable to these two datasets.

Experimental Setup and Evaluation Index

We use the five-fold cross-validation method to start the experiment based on PyTorch version 1.2.0. The division of all datasets is consistent with Ghosh et al. (2020), 20% is used as the test set, 20% is used as the validation set, and 60% is used as the training set. And we use the grid search method

[1]The ASSISTments datasets are retrieved from https://sites.google.com/site/assistmentsdata/home and https://sites.google.com/view/assistmentsdatamining/.

[2]The Statics2011 dataset is retrieved from https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507.

TABLE 2 | The predicted results of different methods on knowledge tracing.

Model	AUC (%)			
	ASSISTments2009	**ASSISTments2015**	**ASSISTments2017**	**Statics2011**
IRT+	77.40*	–	–	–
BKT+	69*	–		75*
DKT	80.53 ± 0.2*	72.52 ± 0.1*	72.63 ± 0.1*	80.20 ± 0.2*
DKVMN	81.57 ± 0.1*	72.68 ± 0.1*	70.73 ± 0.1*	82.84 ± 0.1*
KQN	82.32 ± 0.05*	73.40 ± 0.02*	73.33 ± 0.03*	83.20 ± 0.05*
SAKT	84.8*	85.4*	72.12*	85.3*
AKT-NR	81.69 ± 0.004*	78.28 ± 0.002*	72.82 ± 0.003*	82.65 ± 0.004*
AKT-R	83.46 ± 0.003*	–	77.02 ± 0.002*	–
CAKQN-R	**87.04 ± 0.004**	–	**79.33 ± 0.002**	–
CAKQN-NR	85.54 ± 0.003	**88.88 ± 0.004**	76.45 ± 0.003	**85.43 ± 0.001**

*The symbol * means the result is from other paper. The best results are shown in bold.*

on the validation set to determine the optimal parameters. We use $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, {64, 128, 256, 512}, {64, 128, 256, 512}, {0, 0.05, 0.1, 0.15, 0.2, 0.25}, and {32, 64, 128, 256, 512} as values of the learning rate, the input embedding dimension, the hidden state dimension of LSTM, the dropout rate for the LSTM network, and the dimension of knowledge state vector and question vector, respectively. Finally, we set the maximum number of epochs to 300, the default optimizer to Adam, the learning rate to 10^{-4}, batch size to 32, the input embedding dimension to 128, the dimension of the LSTM hidden layer to 128, the dropout rate to 0.1, the dimension of knowledge state vector and question vector to 128.

With reference to most of the KT research work, we use the area under the curve (AUC) as an evaluation model to predict the performance of the learner's next interaction. The higher the AUC, the better the model's predictive performance.

Experimental Results and Analysis

Comparative Experiment

On four educational datasets, the CAKQN model proposed in our paper is compared with several common traditional network KT model representatives including IRT+ (Pardos and Heffernan, 2011), BKT+ (Yudelson et al., 2013) and neural network representative baseline models, including DKT (Piech et al., 2015), DKVMN (Zhang et al., 2017), KQN (Lee and Yeung, 2019), SAKT (Pandey and Karypis, 2019), AKT (Ghosh et al., 2020), the experimental results are shown in **Table 2**. Note that best models are bold, the results with * are form other paper.

Table 2 lists the performance of all KT methods across all datasets for predicting future learner responses. CAKQN-R and CAKQN-NR represent variants of the CAKQN model with and without the embedding based on the Rasch model, respectively. Similarly, AKT-R and AKT-NR represent variants of the AKT model with and without the embedded Rasch model in Ghosh et al. (2020), respectively. The experimental results show that the CAKQN-R model proposed in this paper is better than the existing model, and its AUC value is 2.945% higher than the existing optimal model AKT-R on average. Note that IRT+ and BKT+ have the lowest prediction performance on the four datasets compared to the neural network representing the four datasets. This indicates that both methods rely on experts to label KC, and the model cannot capture more information like deep neural networks. In the DLKT model, the average prediction performance of the KQN model on the four datasets is significantly improved compared to DKT and DKVMAN. This is because the KQN model is more explanatory in terms of learner knowledge interaction. And CAKQN-R and CAKQN-NR, which also use dot products to represent the interaction process between learner knowledge and questions, have achieved better performance on all datasets. This is related to its different network structure, the monotonic attention mechanism introduced and the embedding based on the Rasch model. Taking a closer look, the SAKT, AKT, and CAKQN models that introduce the attention mechanism and its variants have achieved better results than the general DLKT models such as DKT, DKVMN, and KQN. Because the attention mechanism can link the KC at the next time step with the related KC in the learner's past interaction sequence, the DLKT model with the attention mechanism can more accurately describe the knowledge state of each learner, thereby improving the performance of the model. Among them, the CAKQN-R model achieved better results than other DLKT models with attention mechanisms on the two ASSISTments datasets with question IDs. This proves that the CAKQN-R model can dig more complex features such as forgetting behavior in learner interaction sequences, obtain more accurate learner knowledge status and improve the prediction effect. Comparing the CAKQN-NR and AKT-NR models with the same monotonic attention mechanism, CAKQN-NR model proposed in this paper uses the network structure of LSTM+monotonic attention mechanism to retain the ability of the model to model the sequence, which can not only ensure that the original learner's historical learning interaction sequence is not damaged on the time scale, but also extract complex features of learners such as forgetting behavior. More importantly, it also provides a more interpretable interaction process between learner knowledge and questions, which contributes to a better prediction effect than AKT-R.

TABLE 3 | Experimental comparison between CAKQN and variant that do not use contextual aware question and response representations.

Model	AUC (%)			
	ASSISTments2009	**ASSISTments2015**	**ASSISTments2017**	**Statics2011**
CAKQN^{raw}-NR	84.49 ± 0.004	85.31 ± 0.004	74.84 ± 0.002	85.13 ± 0.001
CAKQN-NR	**85.54 ± 0.003**	**88.88 ± 0.004**	**76.45 ± 0.003**	**85.43 ± 0.001**
CAKQN^{raw}-R	86.12 ± 0.004	–	77.14 ± 0.003	–
CAKQN-R	**87.04 ± 0.004**	–	**79.33 ± 0.002**	–

The best results are shown in bold.

TABLE 4 | Experimental comparison between CAKQN and variants with other attention mechanism.

Model	AUC (%)			
	ASSISTments2009	**ASSISTments2015**	**ASSISTments2017**	**Statics2011**
SAKT	84.8*	85.4*	72.12*	85.3*
CAKQN-NR^{nl}	84.01 ± 0.005	80.52 ± 0.011	71.84 ± 0.004	83.89 ± 0.001
CAKQN-NR	**85.54 ± 0.003**	**88.88 ± 0.004**	**76.45 ± 0.003**	**85.43 ± 0.001**
CAKQN-R^{nl}	85.52 ± 0.004	–	75.44 ± 0.003	–
CAKQN-R	**87.04 ± 0.004**	–	**79.33 ± 0.002**	–

*The symbol * means the result is from other paper. The best results are shown in bold.*

Finally, comparing CAKQN-R and CAKQN-NR, we found that CAKQN-R has better prediction performance on both datasets. This proves that the embedding based on the Rasch model can capture the characteristics of individual differences between different questions of the same KC and the personal abilities of learners, and obtain more accurate personalized representations of learner interaction tuples and questions, thereby improving the performance of the model.

Ablation Experiment

In order to further verify the three key innovations in the CAKQN model: context-aware representation of question vectors, monotonic attention mechanism, and embedding based on the Rasch model, three additional ablation experiments were carried out in this paper. The first experiment is the comparison of CAKQN-R, CAKQN-NR and its variants CAKQN^{raw}-R and CAKQN^{raw}-NR. The structure of CAKQN^{raw}-R and CAKQN^{raw}-NR Question Encoder is the same as the KQN model. It uses a multi-layer perceptron (MLP) to directly input the question embedding to obtain the question vector, the number of hidden layers is 1 and the dimension is 128. The second experiment is to compare CAKQN-R, CAKQN-NR, SAKT models and two variants CAKQN-R^{nl} and CAKQN-NR^{nl} without monotonic attention mechanism. The tow variants use ordinary dot product attention to capture the time dependence in the learner's response data. The last one is the experiment is a comparison between CAKQN-R and variant CAKQN-IRT. The CAKQN-IRT model is based on the DIRT framework proposed in Cheng et al. (2019). Specifically, the *Knowledge State Encoder* and *Question Encoder* components used in the CAKQN-IRT model are the same as CAKQN-R, but the difference is that CAKQN-IRT uses direct embedding instead of Rasch embedding. The *Knowledge State Encoder* component of CAKQN-IRT obtains the learners' ability θ, one *Question Encoder* component inputs the question and KC embedding to obtain the distinction of the question α_j, and the other exactly the same *Question Encoder* component inputs the question embedding to obtain the difficulty of the question β_j. Finally, the obtained parameters are substituted into the two-parameter IRT model formula in section "Traditional Knowledge Tracking Methods" for prediction.

Table 3 shows the results of the first ablation experiment based on the context-aware representation of the question vector. In all datasets, CAKQN-R and CAKQN-NR are better than CAKQN^{raw}-R and CAKQN^{raw}-NR. These results show that our context-aware representation of the question is effective in summarizing the relationship between the question at the next time step and the historical question.

Table 4 shows the results of the second ablation experiment of the monotonic attention mechanism. On all datasets, CAKQN-NR is significantly better than other attention mechanisms, including SAKT. In the case of both using Rasch-based model embedding, CAKQN-R still achieves better results than CAKQN-R^{nl} on the two datasets. The reason for this is that it is different from the common language tasks with strong long-distance dependence between words. The dependence of future learner performance on the past is restricted to a much shorter time window for their forgetting behaviors. Therefore, the monotonic attention mechanism with exponential decay when calculating the attention weight can effectively capture the short-term dependence on the past on the time scale to simulate the forgetting behavior of learners in the learning process.

Table 5 shows the results of the third ablation experiment based on the embedding of the Rasch model. Both models are only tested on the two ASSISTments datasets where the question ID in the dataset is available. On these two datasets, CAKQN-R is significantly better than CAKQN-IRT in the

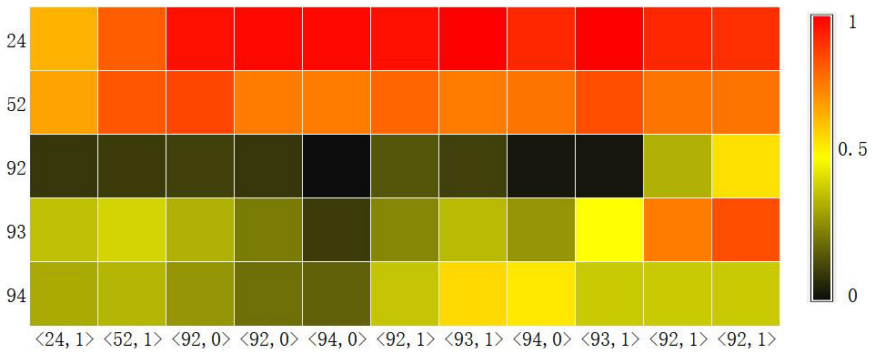


FIGURE 3 | The knowledge level output result of CAKQN-R^{nl} on the ASSISTments2009 dataset.

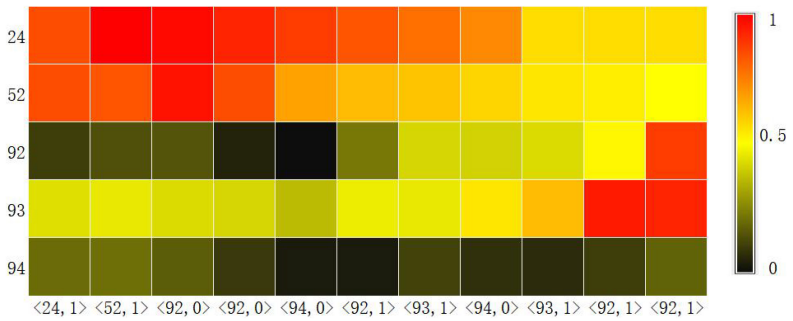


FIGURE 4 | The knowledge level output result of CAKQN-R on the ASSISTments2009 dataset.

TABLE 5 | Experimental comparison between CAKQN and CAKQN-IRT.

Model	AUC (%)	
	ASSISTments2009	**ASSISTments2017**
CAKQN-IRT	84.43 ± 0.015	75.33 ± 0.020
CAKQN-R	**87.04 ± 0.004**	**79.33 ± 0.002**

The best results are shown in bold.

predictive ability of the model. This shows that although CAKQN-IRT incorporates a more complex two-parameter IRT model, CAKQN-R has achieved better results with a simpler model structure. This also confirms that CAKQN-R has more advantages in the knowledge interaction process represented by the dot product calculation in the knowledge query component.

Visualization of Knowledge Tracking

Another basic task of knowledge tracking is to show learners' mastery of each knowledge point in real time. Therefore, we visualized the probability of learners answering correctly at each knowledge point at each time step through the Knowledge Query component. We intercepted the learning records of a learner in the dataset ASSISTments2009 over a period of time, and used the CAKQN-R^{nl} and CAKQN-R model models to track the changes in learners' mastery of 5 knowledge points, as shown in **Figures 3**, **4**. The horizontal axis in the figure represents the interception of the learner's 11 time steps of learning history. The in the tuple represents the learner's KC (knowledge points), represents the learner's answer. The vertical axis represents the 5 knowledge points tracked by the model.

From the visualization results, it can be seen that at the first time step, after the learners answered the exercises containing knowledge points 24 correctly, the tracking results of CAKQN-R^{nl} and CAKQN-R on the learners' knowledge points 24 have been improved (the probability of correct answers increases). The results indicate that the CAKQN-R^{nl} model and the CAKQN-R model will update the mastery of the corresponding knowledge points accordingly after obtaining the learner's historical answer results. In **Figures 3**, **4**, within ten time steps after the learner correctly answered the exercises containing knowledge point 24 at the first time step, CAKQN-R^{nl} did not update the learner's mastery of knowledge point 24, while CAKQN-R showed that the degree of learner's mastery of knowledge point 24 has been declining. It can be seen that the CAKQN-R^{nl} model does not consider the learner's forgetting behavior during the learning period, and the CAKQN-R model fits the learner's actual forgetting behavior during the learning period after introducing the multiplicative exponential decay term. The above results show that both the CAKQN-R model and the CAKQN-R^{nl} model can model the learning process of learners' knowledge status over time. However, the CAKQN-R^{nl} model cannot model the forgetting behavior of learners, while the CAKQN-R model can

model the forgetting behavior of learners, and more accurately track learners' mastery of various knowledge points in real time.

CONCLUSION

Real-time assessment of learners' online learning knowledge level helps to monitor learners' own cognitive status, adjust learning strategies, and improve the quality of online learning. As for four real online education datasets, this paper proposes a CAKQN model based on Rasch model embedding. It uses the vector dot product to describe the interaction process between the learner's knowledge state and the question, and uses the network structure of LSTM + monotonic attention mechanism to capture the question and the learner's personalized contextual representation. Compared with most other knowledge tracking models, it can not only track learners' knowledge status in real time, but also model learners' forgetting behavior.

However, the method presented in this paper has several limitations.

(1) CAKQN uses binary variables to represent the answer to the question as same as other KT methods. This way is not suitable for subjective questions with continuous score distribution. Wang et al. (2017) and Swamy et al. (2018). provide a new way to model subjective questions, they used continuous snapshots of the learner's answers as an indicator of the answer when dealing with learners' programming data. Modeling subjective topics will be the direction of future research.

(2) The adaptive capacity of the model needs to be improved. CAKQN is a supervised training method like other deep knowledge tracking methods, so the predictive ability of the model is dependent on the effect of training on the current dataset. If you are faced with small data sets or other domain datasets, the performance of the model may be poor (Wang Y. et al., 2021).

(3) Like most other KT methods, our method is based on the learner's historical practice record modeling, and involves too few features. In fact, the learning process is very complex, involving many other features such as the text of the question, the learning rate of the student, and the positive/negative emotions that the student generates during the learning process. At present, with the rapid development of technologies such as intelligent perception, wearable devices, and the Internet of Things, multi-modal learning analysis will become a new trend driving intelligent education research (Wang Z. et al., 2021). Under this trend, knowledge tracking will surpass a single behavior modality and gradually develop into a learner model driven by the fusion of multimodal data such as behavior, psychology, and physiology.

DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YC: writing – review and editing, supervision, resources, and conceptualization. GW: methodology, software, validation, and writing – original draft. HZ: visualization and writing – review and editing. PL: data curation. ZC: formal analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported in part by the National Natural Science Foundation of China (62167006 and 61967011), Jiangxi Province Science and Technology Innovation Base Plan-Provincial Key Laboratory Project (20212BCD42001), 03 Special and 5G Projects in Jiangxi Province (20212ABC03A22), National Social Science Fund Key Project (20AXW009), Natural Science Foundation of Jiangxi Province (20202BABL202033 and 20212BAB202017), and Humanities and Social Sciences Key (Major) Project of the Education Department (JD19056). The Jiangxi Province Main Discipline Academic and Technical Leader Training Program–Leading Talent Project (20213BCJL22047).

ACKNOWLEDGMENTS

We acknowledge the financial support provided by all the fundings on this research.

REFERENCES

Bailey, C. D. (1989). Forgetting and the learning curve: a laboratory study. *Manag. Sci.* 35, 340–352. doi: 10.1287/mnsc.35.3.340

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., et al. (2019). "DIRT: deep learning enhanced item response theory for cognitive diagnosis," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, (Gold Coast, AU). doi: 10.1145/3357384.3358070

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User Adapted Interact.* 4, 253–278. doi: 10.1007/BF01099821

Daniluk, M., Rocktäschel, T., Welbl, J., and Riedel, S. (2017). Frustratingly short attention spans in neural language modeling. *arXiv [Preprint]*.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educ. Psychol. Meas.* 58, 357–381. doi: 10.1177/0013164498058003001

Ghosh, A., Heffen, N., and Lan, A. S. (2020). "Context-aware attentive knowledge tracing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (Virtual Event). doi: 10.1145/3394486.3403282

Hu, X., Liu, F., and Bu, C. (2020). Research progress of cognitive tracking models in educational big data. *J. Comput. Res. Dev.* 57, 2523–2546.

Khajah, M. M., Huang, Y., González-Brenes, J. P., Mozer, M. C., and Brusilovsky, P. (2014). "Integrating knowledge tracing and item response theory: a tale of two frameworks," in *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE)*, (Aalborg).

Lee, J., and Yeung, D. Y. (2019). "Knowledge query network for knowledge tracing: how knowledge interacts with skills," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK)*, (Tempe, AZ). doi: 10.1145/3303772.3303786

Liu, T., Wei, C., Liang, C., and Gu, T. (2021). *Research Progress of Knowledge Tracking Based on Deep Learning [OL]*. Available online at: http://kns.cnki.net/kcms/detail/11.1777.TP.20210609.0938.002.html (accessed June 30, 2021).

Pandey, S., and Karypis, G. (2019). "A Self-attentive model for knowledge tracing," in *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*, (Montreal, CA).

Pardos, Z. A., and Heffernan, N. T. (2011). "KT-IDEM: introducing item difficulty to the knowledge tracing model," in *Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization*, (Girona). doi: 10.1007/978-3-642-22362-4_21

Pashler, H., Cepeda, N., Lindsey, R. V., Vul, E., and Mozer, M. C. (2009). Predicting the optimal spacing of study: a multiscale context model of memory. *Adv. Neural inf. Process. Syst.* 22, 1321–1329.

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., et al. (2015). "Deep knowledge tracing," in *Proceedings of the 28th International Conference on Neural Information Processing System (NeurIPS)*, (Cambridge, MA).

Rabiner, L., and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine* 3, 4–16. doi: 10.1109/MASSP.1986.1165342

Sun, J., Li, D., Peng, X., Zou, R., and Wang, P. (2021). Cognitive tracking from a data perspective: framework, problems and enlightenment. *Open Educ. Res.* 27, 99–109.

Swamy, V., Guo, A., Lau, S., Wu, W., Wu, M., Pardos, Z., et al. (2018). "Deep knowledge tracing for free-form student code progression," in *International Conference on Artificial Intelligence in Education*, (Cham: Springer). doi: 10.1007/978-3-319-93846-2_65

Wang, L., Sy, A., Liu, L., and Piech, C. (2017). "Deep knowledge tracing on programming exercises," in *Proceedings of the 4th ACM Conference on Learning*, (New York, NY) doi: 10.1145/3051457.3053985

Wang, Y., Wang, Y., and Zheng, Y. (2021). Multi-modal learning analysis: "Multi-modal"-driven new trends in intelligent education research. *China Audio Visual Educ.* 03, 88–96.

Wang, Z., Xiong, S., Zuo, M., Min, Q., and Ye, J. (2021). Knowledge tracking from the perspective of smart education: status quo, framework and trend. *J. Distance Educ.* 39, 45–54.

Wilson, K. H., Karklin, Y., Han, B., and Ekanadham, C. (2016). Back to the basics: bayesian extensions of IRT outperform neural networks for proficiency estimation. *arXiv [Preprint]*.

Yang, F., and Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Comput. Educ.* 123, 97–108. doi: 10.1016/j.compedu.2018.04.006

Yudelson, M. V., Koedinger, K. R., and Gordon, G. J. (2013). "Individualized bayesian knowledge tracing models," in *International Conference on Artificial Intelligence in Education*, (Berlin). doi: 10.1007/978-3-642-39112-5_18

Zhang, J., Shi, X., King, I., and Yeung, D. Y. (2017). "Dynamic key-value memory networks for knowledge tracing," in *Proceedings of the 26th International Conference on World Wide Web (WWW)*, (Perth, AU). doi: 10.1145/3038912.3052580

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 11 August 2022
DOI 10.3389/feduc.2022.928632

OPEN ACCESS

EDITED BY
Manuel Gentile,
Istituto per le Tecnologie Didattiche ITD—Consiglio Nazionale delle Ricerche, Italy

REVIEWED BY
Wahyu Nanda Eka Saputra,
Ahmad Dahlan University, Indonesia
Alexander Nussbaumer,
Graz University of Technology, Austria

*CORRESPONDENCE
Elizabeth Brooke Cloude
ecloude94@gmail.com

SPECIALTY SECTION
This article was submitted to Digital Learning Innovations, a section of the journal Frontiers in Education

RECEIVED 26 April 2022
ACCEPTED 22 July 2022
PUBLISHED 11 August 2022

CITATION
Cloude EB, Azevedo R, Winne PH, Biswas G and Jang EE (2022) System design for using multimodal trace data in modeling self-regulated learning. *Front. Educ.* 7:928632. doi: 10.3389/feduc.2022.928632

System design for using multimodal trace data in modeling self-regulated learning

Elizabeth Brooke Cloude[1]*, Roger Azevedo[2], Philip H. Winne[3], Gautam Biswas[4] and Eunice E. Jang[5]

[1]Penn Center for Learning Analytics, University of Pennsylvania, Philadelphia, PA, United States, [2]School of Modeling, Simulation, and Training, University of Central Florida, Orlando, FL, United States, [3]Faculty of Education, Simon Fraser University, Burnaby, BC, Canada, [4]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, United States, [5]Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada

Self-regulated learning (SRL) integrates monitoring and controlling of cognitive, affective, metacognitive, and motivational processes during learning in pursuit of goals. Researchers have begun using multimodal data (e.g., concurrent verbalizations, eye movements, on-line behavioral traces, facial expressions, screen recordings of learner-system interactions, and physiological sensors) to investigate triggers and temporal dynamics of SRL and how such data relate to learning and performance. Analyzing and interpreting multimodal data about learners' SRL processes as they work in real-time is conceptually and computationally challenging for researchers. In this paper, we discuss recommendations for building a multimodal learning analytics architecture for advancing research on how researchers or instructors can standardize, process, analyze, recognize and conceptualize (SPARC) multimodal data in the service of understanding learners' real-time SRL and productively intervening learning activities with significant implications for artificial intelligence capabilities. Our overall goals are to (a) advance the science of learning by creating links between multimodal trace data and theoretical models of SRL, and (b) aid researchers or instructors in developing effective instructional interventions to assist learners in developing more productive SRL processes. As initial steps toward these goals, this paper (1) discusses theoretical, conceptual, methodological, and analytical issues researchers or instructors face when using learners' multimodal data generated from emerging technologies; (2) provide an elaboration of theoretical and empirical psychological, cognitive science, and SRL aspects related to the sketch of the visionary system called SPARC that supports analyzing and improving a learner-instructor or learner-researcher setting using multimodal data; and (3) discuss implications for building valid artificial intelligence algorithms constructed from insights gained from researchers and SRL experts, instructors, and learners SRL *via* multimodal trace data.

KEYWORDS

multimodal trace data, self-regulated learning, emerging technologies, system architecture, artificial intelligence

1. Introduction

Technology is woven into the fabric of the twenty-first century. Exacerbated by the pandemic of COVID-19, these emerging technologies have the capacity to increase accessibility, inclusivity, and quality of education across the globe (UNESCO, 2017). Emerging technologies include serious games, immersive virtual environments, simulations, and intelligent tutoring systems that have assisted learners in developing self-regulated learning (SRL) and problem-solving skills (Azevedo et al., 2019) across multiple domains (Biswas et al., 2016; Azevedo et al., 2018; Winne, 2018a; Lajoie et al., 2021), populations, languages, and cultures (Chango et al., 2021). Empirical evidence shows that SRL with emerging technology results in better learning gains compared to conventional methods (Azevedo et al., 2022). These technology-rich learning environments can record learners' multimodal trace data (e.g., logfiles, concurrent verbalizations, eye movements, facial expressions, screen recordings of learner-system interactions, and physiological signals) that instructors and education researchers can use to systematically monitor, analyze, and model SRL processes, and study their interactions with other latent constructs and performance with overall goals to augment teaching and learning (Azevedo and Gašević, 2019; Hadwin, 2021; Reimann, 2021).

Emerging evidence points to key roles that multimodal data can play in this context (Jang et al., 2017; Taub et al., 2021) and has sparked promising data-driven techniques for discovering insights into SRL processes (Cloude et al., 2021a; Wiedbusch et al., 2021). Yet, major issues remain regarding roles for various SRL processes (e.g., cognitive and metacognitive; Mayer, 2019) and their properties: evolution or recursive nature over time, frequency and duration, interdependence, quantity vs. quality (e.g., accuracy in metacognitive monitoring), and methods for fusing multimodal trace data to link SRL processes to learning task performance. As research using multimodal trace data unfolds, we perceive an increased need to understand how instructional decisions can be forged by modeling regulatory patterns reflected by multimodal data in both learners and their instructors. We pose a fundamental question: Has the field developed the knowledge and the supporting processes to help researchers and instructors interpret and exploit multimodal data to design productive and effective instructional decisions?

In this paper, we provide an elaboration on psychological aspects related to the design of a teaching and learning architecture called SPARC that allows researchers or instructors to standardize, process, analyze, recognize, and conceptualize (SPARC) SRL signals from multimodal data. The goal is to help researchers or instructors represent and strive to understand learners' real-time SRL processes, with the aim to intervene and support ongoing learning activities. We envision a SPARC system to reach this goal. Specifically, we recommend that the design of SPARC should embody a framework grounded in (1) conceptual and theoretical models of SRL (e.g., Winne, 2018a); (2) methodological approaches to measuring, processing, and modeling SRL using real-time multimodal data (Molenaar and Järvelä, 2014; Segedy et al., 2015; Bernacki, 2018; Azevedo and Gašević, 2019; Winne, 2019), and (3) analytical approaches that coalesce etic (researchers/instructors) and emic (learners) trace data to achieve optimal instructional support. We first discuss previous studies using learners' multimodal trace data to measure SRL during learning activities with emerging technologies. Next, we describe challenges in using these data to capture, analyze, and understand SRL by considering recent developments in analytical tools designed to handle challenges associated with multimodal learning analytics. Lastly, we recommend a hierarchical learning analytics framework and discuss theoretical and empirical guidelines for designing a system architecture that measures (1) learners' SRL alongside (2) researchers'/instructors' monitoring, analyzing, and understanding of learners' SRL grounded in multimodal data to forge instructional decisions. Implications of this research could pave the way for training artificial intelligence (AI) using data insights gained from researchers, instructors, and experts within the field of SRL that vary by individual characteristics including training background/experience, country, culture, gender, and many other diversity aspects. Algorithms trained using data collected on a diverse sample of interdisciplinary and international (1) SRL experts and researchers, (2) instructors, and (3) learners has the potential to automatically detect and classify SRL constructs across a range of data channels and modalities could serve to mitigate the extensive challenges associated with using multimodal data and assist educators in making effective instructional decisions guided by both theory and empirical evidence.

1.1. Characteristics of multimodal data used to reflect SRL

To gather multimodal data about SRL processes during learning, learners are instrumented with multiple sensors. Examples include electro-dermal bracelets (Lane and D'Mello, 2019), eye tracking devices (Rajendran et al., 2018), and face tracking cameras to capture facial expressions representing emotions (Taub et al., 2021). These channels may be supplemented by concurrent think-aloud data (Greene et al., 2018), online behavioral traces of learners using features in a software interface (Winne et al., 2019) and gestures and body movements in an immersive virtual environment (Raca and Dillenbourg, 2014; Johnson-Glenberg, 2018). This wide array of data can reflect when, what, how, and how long learners interact with specific elements in a

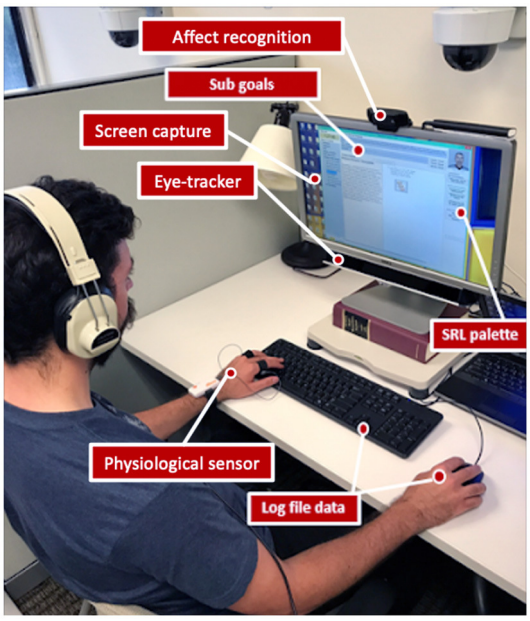


FIGURE 1
Experimental set-up to collect multimodal SRL data with MetaTutor.

learning environment—e.g., reading and highlighting specific text, inspecting diagrams, annotating particular content, manipulating variables in simulations, recording, and analyzing data in problem-solving tasks, and interacting with pedagogical agents (Azevedo et al., 2018; Winne, 2019). Multimodal data gathered across these channels offer advantages in representing latent cognitive, affective, metacognitive, and motivational processes that are otherwise weakly signaled in any single data channel (Greene and Azevedo, 2010; Azevedo et al., 2018) (Figure 4).

A typical laboratory experimental set-up shown in Figure 1 illustrates a college student instrumented during learning with MetaTutor, a hypermedia-based intelligent tutoring system designed to teach about the human circulatory system (Azevedo et al., 2018). In addition to pre- and post-measures of achievement and self-report questionnaires not represented in the figure, multimodal instrumentation gathers a wide range of data about learning and SRL processes. Mouse-click data indicate when, how long, and how often the learner selects a page to study. Features and tools available for to the learner in a palette, such as self-quizzing and typing a summary, identify when the learner makes metacognitive judgments about knowledge (Azevedo et al., 2018) and how that might change across different learning goals (Cloude et al., 2021b). An electro-dermal bracelet records signals documenting changes in skin conductance produced by sympathetic innervation of sweat glands, a signal for arousal that can be matched to the presence of external sensory stimuli (Lane and D'Mello, 2019; Messi and Adrario, 2021; Dindar et al., 2022). Eye movements operationalize what, when, where, and how long the learner attends to, scans, revisits, and reads (or rereads) content and consults displays, such as a meter showing progress toward goals (Taub and Azevedo, 2019; Cloude et al., 2020). Dialogue recorded between the learner and any of the four pedagogical agents embedded in MetaTutor identify system-provided scaffolding and feedback. Screen-capture software records and time stamps how and for how long the learner interacts with all these components and provides valuable contextual information supplementing multimodal data. A webcam samples facial features used to map the sequence, duration, and transitions between affective states (e.g., anger, joy) and learning-centered emotions (e.g., confusion).

Figure 2 illustrates examples of multimodal data used to study SRL across several emerging technologies including MetaTutor, Crystal Island, and MetaTutor IVH (Azevedo et al., 2019). The figure omits motivational beliefs because motivation has been measured almost exclusively using self-reports (Renninger and Hidi, 2019). Multimodal data structures are wide in scope, complexly structured and richly textured. For example, a learner reading about the anaphase stage of cell division may have metacognitively elected to apply particular cognitive tactics (e.g., selecting key information while reading, then assembling those selections across the text and diagram as a summary). At that point, eye-gaze data show repeated saccades and fixations between text and diagrams as the learner utters a metacognitive judgment captured *via* think-aloud, "I do not understand the structures of the heart presented in the diagram." Concurrently, physiological data reveal a spike in heart rate and analysis of the learner's facial expressions indicate frustration. Inspecting and interpreting this array of time-stamped data sampled across multiple scales of measurement and spanning several durations pose significant challenges for modeling cognition, affect, metacognition, and motivation. Which data channels relate to the different SRL features (cognition, affect, metacognition, and motivation)? Is one channel better at operationalizing a specific SRL feature? How should the different data channels be configured so that researchers can accurately monitor, analyze, and interpret SRL processes in real-time? What is (are) the appropriate temporal interval(s) for sampling each data channel, and how are characterizations across data channels used to support accurate and valid interpretations of latent SRL processes? Assuming these questions are answered, how can researchers be guided to make instructional decisions that support and enhance learners' SRL processes? We suggest guidelines to address these questions in the form of a SPARC system. Our paper is based in theoretical and empirical literature from the science of learning, and evolving understanding about multimodal trace data.

FIGURE 2
Examples of specific types of multimodal data to investigate cognitive, affective, and meta-cognitive SRL process with different emerging technologies (Azevedo et al., 2019).

1.2. Challenges in representing SRL using multimodal data

Time is a necessary yet perplexing feature needing careful attention in analyses of multimodal data sampled over multiple channels. How should data with differing frequencies be synchronized and aligned when modeling processes? To blend multisynchronic data, time samples need to be rescaled to a uniform metric (e.g., minutes or seconds). Multimodal data may require filtering to dampen noise and lessen measurement errors. Decisions about these adjustments can be made usually only after learners have completed segments in or an entire study session. Judgments demand intense vigilance as researchers and instructors scan multimodal data and update interpretations grounded in multimodal data. If researchers or instructors attempt to monitor and process multimodal data in real-time to intervene during learning—e.g., prompting learners to avoid or correct unproductive studying tactics—vigilance will be one key.

In the presence of dense and high-velocity data, critical signals in multimodal data that should steer instructional decision-making may be missed as demonstrated in Claypoole et al.'s (2019) study. Their findings showed increases in stimuli per minute decreased participants' sensitivity (discriminating hits and false alarms) and increased time needed to detect pivotal details (Claypoole et al., 2019). As well, because vigilance declines over time and tasks (Hancock, 2013; Greenlee et al., 2019), counters need to be developed if multimodal data are to be useful inputs for real-time instructional decision-making to support learners' SRL. Furthermore, a particular and pressing challenge for moving this research forward is determining what information can be used from learners, such as *who* will be allowed to access potentially personal data, and *how* might such users obtain permissions to ethically use the data (Ifenthaler and Schumacher, 2019), meanwhile maintaining confidentiality, reliability, security, privacy, among many others that align with security and privacy policies that may vary across international lines (Ifenthaler and Tracey, 2016).

International researchers have begun to engineer systems to manage challenges associated with processing, analyzing, and understanding multimodal data. For example, SensePath (Nguyen et al., 2015) was built in the United Kingdom and designed to reduce demand for vigilance by providing visual tools that support articulating multichannel qualitative information unfolding in real-time, such as transcribed audio mapped onto video recordings. Blascheck et al. (2016) developed a similar visual-analytics tool in Germany to support coding and aligning mixed-method multimodal data gathered over a learning session in the form of video and audio recordings, eye-gaze tracks, and behavioral-interactions. Their system was designed to support researchers in (1) identifying patterns, (2) annotating higher-level codes, (3) monitoring data quality for errors, and (4) visually juxtaposing codes across researchers to foster discussions and contribute to inter-rater reliability (Blascheck et al., 2016). These systems illustrate progress in engineering tools researchers and instructors need to work with complex multimodal data, such as those required to reflect learning and SRL. But two gaps need filling. First, systems developed so far could further mine and apply research on how humans make sense of information derived from complex multimodal data. Second, systems have not yet been equipped to gather and mine data about how researchers and instructors use system features using data representations and visual tools. Furthermore, how might researchers and instructors use system features differently as their goals and intentions, training, and beliefs about phenomenon may vary? In other words, developing models that represent how diverse users leverage the system and its features need to be considered in future work to build a multi-angled view of the total system.

One notable system designed for multimodal signal processing and pattern recognition in real-time is the Social Signal Interpretation framework (SSI) developed in Germany by Wagner et al. (2013). SSI was engineered to simultaneously process data ranging from physiological sensors and video recordings to Microsoft's Kinect. A machine-learning (ML) pipeline automatically aligns, processes, and filters multimodal data in real-time as it is collected. Once data are processed, automated recognition routines detect and classify learners' activities (Wagner et al., 2013). Another multimodal data tool, SLAM-KIT (Noroozi et al., 2019), was built in the Netherlands and designed to study SRL in collaborative contexts. It reduced the volume and variety in multimodal data to allow teachers, researchers, or learners to easily navigate in and across data streams, analyze key features of learners' engagements, and annotate and visualize variables or processes that analysts identify as signals of SRL. Notwithstanding the advances these systems represent, issues remain. One is how to coordinate (a) data across multiple channels with multiple metrics alongside (b) static and unfolding contextual features upon which learners pivot when they regulate their learning (Kabudi et al., 2021). Another target for improvement is supporting researchers to monitor, analyze and accurately interpret matrices of the multimodal data for tracking SRL processes. Factors that may affect such interpretations include choosing and perhaps varying optimal rates to sample data, synchronizing and temporally aligning data in forms that support searching for patterns, and articulating online data with contextual data describing tasks, domains learners study, and characteristics of settings that differentiate the lab from the classroom from home, and individual vs. collaborative work. All these issues have bearing on opportunities to test theoretical models (e.g., Winne, 2018a) and positively influence learning.

1.3. Overview of the SPARC architecture

Making effective just-in-time and just-in-case instructional decisions demands expertise in monitoring, analyzing, and modeling SRL processes. Are researchers and/or instructors equipped to meet these challenges when delivered fast-evolving multimodal data? To address these issues, we discuss a hierarchical learning analytics framework and guidelines for designing a theoretically and empirically-based suite of analytical tools to help researchers and/or instructors solve challenges associated with receiving real-time multimodal data, monitoring SRL, assembling interventions, and tracking dynamically unfolding trajectories as learners work in emerging technologies. The SPARC system we recommend is a dynamic data processing framework in which multimodal data are generated by (1) learners, (2) researchers and/or instructors, and (3) the system itself. These streams of data are automatically[1] processed in real-time negative feedback loops. Two features

1 We cannot elaborate on this process due to space limitations. However, there are several tools currently used by interdisciplinary

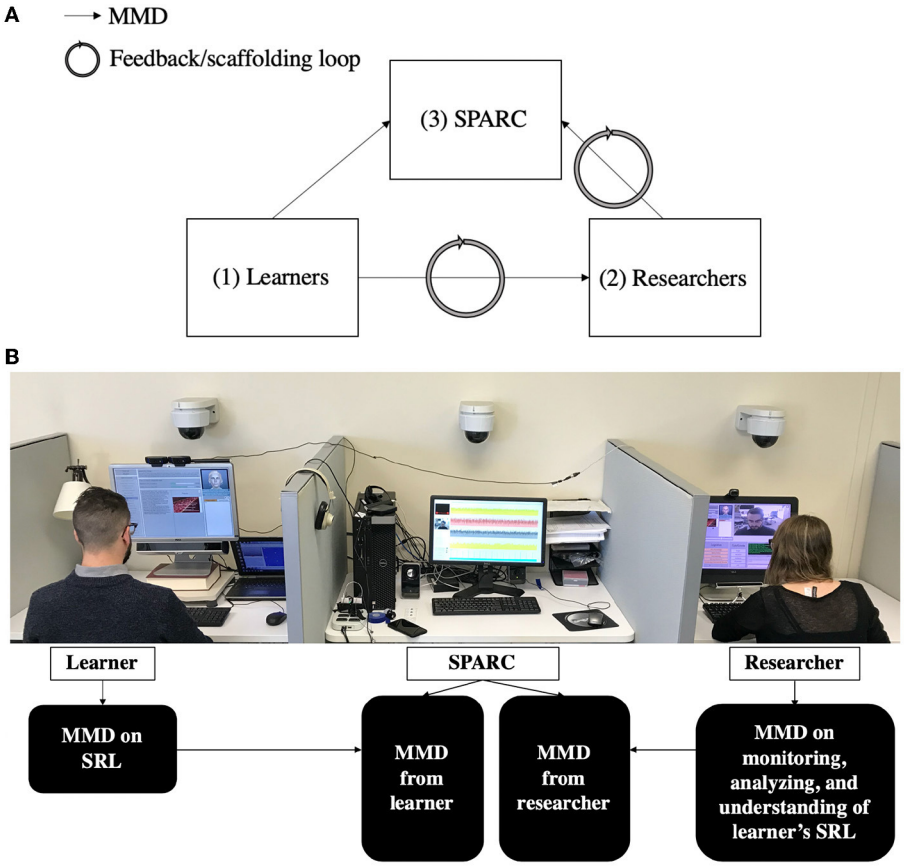


FIGURE 3
(A) SPARC capturing the learner's and researcher's multimodal data; **(B)** Hierarchical architecture for data processing and feedback/scaffolding loops for both (1) learners and (2) researchers; MMD, multimodal data.

distinguish SPARC from other tools. First, researchers and/or instructors are positioned in three roles: data generators, data processors, and instructional decision makers. Second, the overall system has a triifb-partite structure designed to record series of multimodal data for real-time processing across the timeline of instructional episodes. The target SPARC aims for is iteratively tuning data gathering, data processing, and scaffolding for both learners and researchers/instructors, thereby helping both players more productively self-regulate their respective and interactive engagements (see Figure 3).

Imagine an instructor and learner are about the engage in a learning session with an emerging technology. See Figure 3–both the users (researchers and learners) interact with content while their data are recorded on such interaction. For example, both instructor and learner are instrumented with multiple sensors, including a high-resolution eye tracker and physiological bracelet, meanwhile, both users' video, audio, and screens are being recorded. Further, once the learner begins interacting with content, their data are recorded and SRL variables are generated in real-time. These data are displayed to the instructor so they can see what the learner is doing such as where their eyes are attending to specific text and/or diagrams, including the sequence and amount of time they are engaging with content. The instructor can also see the learner's physiology spikes, facial expressions of emotions, screen recording, and speech. Meanwhile, the instructor is also being recorded with sensors. Once the learner begins engaging with materials, data on the instructor measure the degree to which they attend—i.e., monitor, analyze, and understand the learner's SRL *via* data channels and modalities over the learning activity. From these data, SPARC can calculate the degree to which the instructor is biased to oversample a specific channel of learner data, say, eye-gaze behaviors. For instance, SPARC can detect this bias when the instructor's eye-gaze and logfiles data show they infrequently sample other data channels carrying critical SRL signals. Here, SPARC should take three steps. First, alert the instructor to shift attention, e.g., by posting a notification, "*Is*

researchers to view, process, and analyze learners' multimodal data in real-time such as the iMotions's research platform.

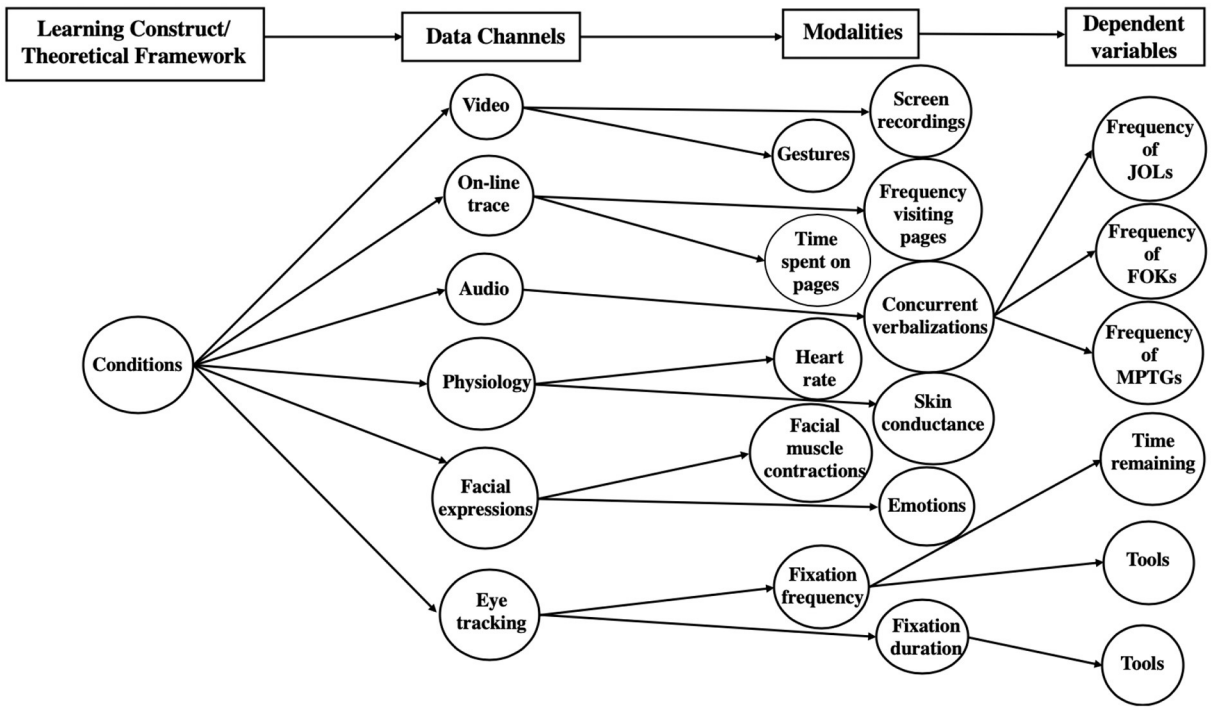


FIGURE 4
Pipeline architecture for capturing conditions using Winne's model. JOL, judgment of learning; FOK, feeling of knowing; MPTG, monitoring progress toward goals.

variance in the learner's eye-gaze data indicating a change in standards used for metacognitive monitoring?." Second, SPARC varies illumination levels of its panels to cue the instructor to shift attention to the panel displaying learner eye movements. Third, a pop-up panel shows the instructor a menu of alternative interventions. In this panel, each intervention is described using a 4-spoke radar chart grounded in prior data gathered from other learners: the probability of learner uptake e.g., Bayesian knowledge tracing (Hawkins et al., 2014), the cognitive load associated with the intervention, negative impact on other study tactics such as note-taking, and learner frustration triggered upon receiving SPARC's recommendation to adapt standards for metacognitively monitoring understanding. Then, SPARC monitors the instructor's inspection of display elements to update its model of the instructor's biases for particular learner variables—e.g., a preference to limit learner frustration when selecting an intervention to be suggested to the learner. And later, as SPARC assembles data about how the learner reacts to the instructor's chosen intervention, the model of this learner is updated to sharpen a forecast about the probability of intervention uptake and impact of the instructor's chosen intervention on the profile of study tactics this learner uses. SPARC'S complex and hierarchical approach to recording, analyzing, and interacting with both agents in instructional decision-making and self-updating models both pushes SPARC past the boundaries of other multimodal systems. By dynamically updating models of learners and instructors (or researchers), instructional decisions are grounded and iteratively better grounded on the history of all three players—self-regulating learners, self-regulating instructors, and the self-regulating system itself. If widely distributed to create genuinely diverse big data, SPARC would significantly advance research in learning science and mobilize research based on expanding empirical evidence about SRL and interventions that affect it (e.g., Azevedo and Gašević, 2019; Winne, 2019).

1.3.1. Information processing theory of self-regulated learning

The first step toward a SPARC system is building a pipeline architecture that operationalizes a potent theoretical framework to identify, operationalize, and estimate values for parameters of key variables (Figure 4). From a computer-science perspective, a pipeline architecture captures a sequence of processing and analysis routines such that outputs of one routine (e.g., capturing multimodal data on researchers and learners) can be fed directly into the next routine without human intervention (e.g., processing learners' and researchers'

multimodal data separately). Overall, an ideal system should support valid interpretations of latent causal constructs. SPARC adopts an information processing view of self-regulated learning along with assumptions fundamental to this perspective (Winne and Hadwin, 1998; Winne, 2018a, 2019).

According to the Winne-Hadwin model of SRL, human learning is an agentic, cyclic, and multi-faceted process centered on monitoring and regulating information in a context of physical and internal conditions bearing on cognitive, affective, metacognitive, and motivational (CAMM) processes during learning (Malmberg et al., 2017; Azevedo et al., 2018; Schunk and Greene, 2018; Winne, 2018a). Individual differences, such as prior knowledge about a domain and self-efficacy for a particular task, and contextual resources (e.g., tools available in a learning environment) set the stage for a cycle of learning activity (Winne, 2018a). Consequently, to fully represent learning as SRL requires gathering data to represent cognitive, affective, metacognitive, and motivational processes while learners (and researchers) learn, reason, problem solve, and perform. Also, to ensure that just-in-case instructional decisions can be grounded in this dynamic process and assumptions of SRL (Winne, 2018a), we propose capturing multimodal data about the instructional decision maker (i.e., researcher) is just as relevant and important as capturing multimodal data about the learner. Thus, the pipeline architecture should be fed data across channels and modalities tapping cognitive, affective, metacognitive, and motivational processes separately for both learners and researchers (see Figure 3). The Winne-Hadwin model of agentic SRL (Winne and Hadwin, 1998; Winne, 2018a) describes learning in terms of four interconnected and potentially nonsequential phases. In Phase 1, the learner surveys the task environment to identify internal and external conditions perceived to have bearing on the task. Often, this will include explicit instructional objectives set by an instructor. In phase 2, based on the learner's current (or updated) understanding of the task environment, the learner sets goals and develops plans to approach them. In phase 3, tactics and strategies set out in the plan are enacted and features of execution are monitored. Primary among these features is progress toward goals and subgoals the learner framed in phase 2; and emergent characteristics of carrying out the plan, such as effort spent, pace, or progress. In phase 3, the learner may make minor adaptations as judged appropriate. In phase 4, which is optional, the learner reviews work on the task writ large. This may lead to adaptive re-engagement with any of the preceding phases as well as forward reaching transfer (Salomon and Perkins, 1989) to shape SRL in similar future tasks.

For the physical set-up[2] illustrated in Figure 3, entries in Table 1 demonstrate a complex coordination between learners' and researchers' multimodal data facilitated by a SPARC system. In this table, we provide two examples that map assumptions based on Winne's phases to the learners' and researchers' multimodal SRL data (Winne, 2018a). Included are a researcher's monitoring, analyzing, and understanding of a learner's SRL based on the learner's multimodal data and instructional interventions arising from the researcher's inferences about the learner's SRL. Two contrasting cases are provided. The first is a straightforward example of a learner's multimodal data that is easy to monitor, analyze, and understand. This leads, subsequently, to an accurate inference about SRL by the researcher who does not require SPARC to intervene in supporting the researcher's instructional intervention. A second scenario is more complex. The learner presents several signals in multimodal data, which could reflect multiple and diverse issues related to their motivation, affect, and cognition. The researcher must intervene to scaffold and prompt the learner but it is not clear where to begin given multiple instructional concerns. So, SPARC intervenes to scaffold the researcher to optimize instructional decision-making based on pooled knowledge about the learner's SRL and the researcher's past successful interventions.

Throughout each phase of SRL, five facets characterize information and metacognitive events, encapsulated in the acronym COPES. Conditions are resources and constraints affecting learning. Time available to complete the task, interest, and free or restricted access to just-in-time information resources are examples. Operations are cognitive processes learners choose for manipulating information as they address the task. Winne models five processes: searching, monitoring, assembling, rehearsing, and translating (SMART). Products refer to information developed by operations. These may include knowledge recalled, inferences constructed to build comprehension, judgments of learning, and recognition of an arising affect. Evaluations characterize the degree to which products match standards, criteria the learner set or adopted from external sources (e.g., an avatar) to operationalize success in work on the task (e.g., pace), and its results. Since SPARC's pipeline architecture is intended to capture learner's SRL processes, we argue it is critical to investigate how to map multimodal data from specific data channels to the theoretically-referenced constructs in the Winne and Hadwin model of SRL and the cognate models for tasks (COPES) and operations (SMART) within tasks. For instance, what data channels or combinations of data channels and modalities best represent a cognitive strategy? Do these data also indicate elements of metacognition? A system such as SPARC should help answer these questions by examining how researchers' and learners' multimodal data might reflect these processes and how those processes impact performance and instructional decision-making, respectively.

We outline a potentially useful start for mapping data channels and modalities to theoretical constructs based on

2 This figure is for illustration purposes only since, ideally, we would physically separate the researcher and learner to avoid bias, social desirability, etc.

TABLE 1 Learner's and researcher's multimodal SRL data aligned with phase 2 of Winne's model of SRL and corresponding instructional strategies based on umambiguous signals in data.

Learning context	Learner's MMD	Researcher's MMD	SPARC
1. Learner engages with biology content and sets • Goals and • Plans based on their current (or updated) understanding 2. Researcher observes learner's MMD on their IF 3. SPARC observes both the researcher's and learner's MMD	• Concurrent verbalizations *via* audio recording (e.g., *"my goal is to learn about how blood flows through the four chambers of the heart."*; • Screen recordings showing learner-system interactions; • Eye gaze illustrating where learner is searching for relevant information on the IF	• Concurrent verbalizations *via* audio recording (e.g., *"I will review learner's data within the context of their goal.... They are not monitoring the right information."*; • Screen recordings of system interactions with learner's MMD; • Eye gaze on learner's interaction with content on IF	SPARC observes and updates its user model based on 1. Researcher's MMD, including eye gaze, utterances, and screen recordings of researcher-system interactions 2. Learner's MMD, including eye gaze, utterances, and screen recordings of learner-system interactions

the COPES model (Winne, 2018a, 2019; Winne P., 2018), specific to conditions[3] (Figure 3). As specified in Figure 3, some criterion variables refer to information captured as a learner verbalizes monitoring of engagement in a task (e.g., frequency of judgment of learning, feeling of knowing). Other data obtained from eye tracking instrumentation represents learners' assessing conditions, such as time left for completing the task signaled by viewing a countdown timer in the interface. These data expand information on conditions beyond records of how frequently learners visit pages in MetaTutor (Azevedo et al., 2018), edit a causal map in Betty's Brain (Biswas et al., 2016), or highlight text in nStudy (Winne et al., 2019). Together, these multimodal data characterize how, when, and with what the learner is proceeding with the task and engaging in SRL. A pipeline architecture for SPARC affords modularity as illustrated in Figure 3. The pipeline can customize data cleaning, pre-processing, and analysis routines for each one of the sensing modalities (e.g., think-alouds, eye-gaze and on-line behavioral traces). It also provides separate analysis routines for each of the constructs (e.g., conditions vs. operations vs. products), data channels, modalities, and criterion variables of which can work with the time series (or event sequence data) generated by the previous module. In sum, supporting researchers in constructing meaningful and valid inferences about SRL from multimodal signals in learners' data requires building a pipeline architecture aligned to a theoretical model of SRL. But this begs a key question. After variables are mapped onto a model of SRL, how can researchers' inferences be reasonably adjudicated? How valid are they?

3 Due to space limitations, we will not go into depth on how the SPARC pipeline will be structured to capture, operationalize, and process variables across settings, tasks, and domains that are aligned with the information-processing theory of SRL, including COPES and SMART.

2. Methods

2.1. Empirical synthesis on monitoring, analyzing, and understanding of multimodal data

Setting aside for the moment issues of alignment between SPARC and the Winne and Hadwin model of SRL, it is prudent to synthesize empirical research related to what, when, and how researchers might examine learners' multimodal data to model and understand dynamically unfolding SRL processes. Consequently, we next examine research on humans (a) monitoring information for patterns, (b) analyzing signals detected in patterns, and (c) constructing understanding(s) of this information matrix by monitoring and analyzing stimuli. Further, we emphasize previous methods and findings in literature as potential directions for leveraging trace data to define cognitive and metacognitive aspects of SRL constructs such as monitoring, analyzing, and understanding of SRL in researchers, instructors, and learners. Finally, we discuss challenges and future directions for the field to consider in ways to leverage multimodal data to advance the design of emerging technologies in modeling SRL.

2.1.1. Monitoring real-time multimodal data

Cognitive psychological research on information processing and visual perception—specifically, selective attention, and bottom-up/top-down attentional mechanisms (Desimone and Duncan, 1995; Desimone, 1996; Duncan and Nimmo-Smith, 1996)—is a fruitful starting point to examine factors that bear on how approaches for defining monitoring of SRL signals in multimodal data. When instructors or researchers encounter multimodal data, only a select partition of the full information matrix can be attended to at a time. One factor governing what

can be inspected is the size of the retina which determines how much visual information is available for processing. Where humans look typically reveals foci of attention and, thus, what information is available for processing (van Zoest et al., 2017). Multimodal data are typically presented across multiple displays and, often, as temporal streams of data. Attending to displays, each representing a particular modality, precludes attending to other data channels. This gives rise to two key questions: (1) What is selected? (2) What is screened out? (Desimone, 1996). One theory describing a mechanism for controlling attention is the biased-competition theory of selective attention (Desimone, 1996; Duncan and Nimmo-Smith, 1996). It proposes a biasing system driven by bottom-up and top-down attentional control. Bottom-up attentional control is driven by stimuli, e.g., peaks in an otherwise relatively flat progression of values in the learners' data that SPARC supplies to researchers and instructors. Bottom-up visual attention is skewed to sample information in displays based on shapes, sizes, and colors, and motion, while top-down attentional control is influenced by a researcher's or instructors goals and knowledge—declarative, episodic, and procedural—both of which are moderated by their beliefs and attitudes (Anderson and Yantis, 2013; Anderson, 2016). In the context of multimodal data SPARC displays, attention is directed in part by a researcher's knowledge about data in a particular channel, e.g., the relative predictive validity of facial expressions compared to physiological signals as indicators of learners' arousal. Another factor affecting the researcher's or instructors attention is the degree of training or expertise in drawing grounded inferences about an aspect of learners' SRL—e.g., recognizing facial expressions of frustration. A third factor is the researchers' or instructors preferred model of learning (e.g., this is what I believe SRL looks like). In the case of SPARC, this is familiarity with and commitment to the 4-phase model of SRL and the COPES schema within each phase. Thus, a key aspect of designing a system like SPARC required situating multimodal data around the goal of the (a) session (e.g., detect SRL in a learner's multimodal data) and (b) the user's goals, beliefs, training, education, and familiarity with and commitment to the 4-phase model of SRL and COPES schema within each phase. Variables defining monitoring behaviors need to be contextualized or evaluated against these criteria or set of standards.

2.1.2. Data channels that capture monitoring behaviors

Eye-tracking methodologies have opened a window into capturing implicit monitoring processes (Scheiter and Eitel, 2017). Mudrick et al. (2019) studied pairs of fixations to identify implicit metacognitive processing. Participants' fixations across text and a diagram were examined for dyads where the information was experimentally manipulated to be consistent or inconsistent (e.g., the text described blood flow but a diagram illustrated lung gas exchange). For each dyad, participants metacognitively judged how relevant information in one medium was to information in the other medium. When information was consistent across dyads, participants more frequently traversed sources and made more accurate metacognitive judgments on the relevance of information in each medium. Eye-gaze data were a strong indicator of metacognitive monitoring and accuracy of judgments. Eye-gaze data also signal other properties of metacognition. Participants in Franco-Watkins et al.'s (2016) study were required to make a decision in a context of relatively little information. In this case, they fixated longer on fewer varieties of information. As variety of information increased, fixations settled on more topics for shorter periods of time. Variety and density of information affected metacognitive choices about sampling information in their complex information displays (Franco-Watkins et al., 2016).

These findings forecast how researchers or instructors may attend to multimodal data with SPARC. For example, if less information is available—i.e., a learner is not thinking aloud and displays a facial expression signaling confusion, will researchers or instructors bias sampling of data in classifying the learner's state by fixating longer on a panel displaying facial expression data, or will they suspend classification to seek data in another channel? A SPARC system would need to collect information on if, when, where, and for how long the user attends to a specific modality or channel, and then prompt the researcher or instructors to introduce data from another channel before classifying learner behavior and recommending a shift in learner behavior. The value of eye-gaze data as proxies for implicit processes such as attention and metacognitive monitoring lead us to suggest that SPARC measures researchers' or instructors' eye-gaze behaviors. Sequences of saccades, fixations, and regressions while monitoring multi-panel displays of learners' multimodal SRL data during a learning activity may reveal how, when, and what researchers or instructors are monitoring in the learners' multimodal data as they strive to synthesize information across modalities. Furthermore, information on what the user is attending to would reveal what the user is *not* attending to that may be potentially important. It would be important for SPARC to also define lack of attention to potentially operationalize the users' goal or intention and whether they are aligned with detecting SRL processes across the data. These data can track whether, when, and for how long users attend to discriminating or non-discriminating signals, sequences, and patterns in learners' multimodal data. Logged across learners and over study sessions, SPARC's data could be mined to model a researcher's or instructors' biases for particular channels in particular learning situations. Beyond eye-tracking data, can other methodologies reveal how researchers analyze learners' engagements during a learning session?

2.2. Analyzing real-time multimodal data

Analyzing and reasoning are complex forms of cognition (Laird, 2012). They dynamically combine knowledge and critical-thinking skills such as inductive and deductive reasoning and may involve episodically-encoded experiences (Blanchette and Richards, 2010). Theoretically, after researchers or instructors allocate attention to multimodal data, they must analyze and then reason about patterns and their sequences in relation to SRL phases and processes. As such, SPARC needs to measure how researchers or instructors search and exploit patterns of multimodal data (relative to other patterns) to make inferences about learners' SRL and recommend interventions. A fundamental issue here is to operationally define a pattern in a manner that achieves consensus among researchers or instructors and can be reliably identified when multimodal data range across data channels. SPARC's capabilities should address questions such as is there a pattern in eye-gaze data that is indicative of SRL? What patterns of gaze data does a researcher use to infer a learner's use and adaptation of tactics, or occurrences of metacognitive monitoring prompted by changing task conditions? In what ways do researchers' or instructors' eye-gaze patterns change over time (e.g., are they focusing on one data channel or more than one? Does their degree of attention change to other data channels?) and, if paired with other modalities, does this change reflect limits or key features in learners' SRL? Do changes in one modality of data indicate changes in other multimodal data and are these changes related to a user's instructional decision-making? We establish a ground truth regarding the validity and reliability of learners', instructors', and researchers' multimodal data patterns (see Winne, 2020). SPARC should be capable of detecting when users are analyzing and reasoning based on multimodal data describing how the user examined learners' multimodal eye-gaze behavior, interactions with the content (logfiles), physiology profile, facial expressions, and other channels. Again, researchers' or instructors' multimodal data play a key role in successively tuning the overall system.

2.2.1. Data channels that capture analyzing behaviors

Some research capturing data to infer implicit processing, such as analyzing and reasoning, used online behavioral traces (Spires et al., 2011; Kinnebrew et al., 2015; Taub et al., 2016); other studies used eye movements data (Catrysse et al., 2018) or concurrent think-aloud verbalizations (Greene et al., 2018). Taub et al. (2017) analyzed learners' clickstream behavior as relevant or irrelevant to the learning objective (e.g., learn about biology), and then applied sequential pattern mining analyses. Two distinct patterns of reasoning differed in efficiency, defined as fewer attempts toward successfully meeting the objective of the learning session. Defining logged learner actions based on relevance to a learning objective is useful for capturing and measuring analyzing and reasoning behaviors (Taub et al., 2017). This technique could also be applied to the researchers' multimodal data as well. For example, do the researchers' mouse clicks, keyboard strokes, etc. reflect their analysis of the learners' multimodal data in relation to meeting the objective of the learning session—e.g., learning about the circulatory system. Is the researcher selecting modalities to evaluate whether the learner is working toward this objective (e.g., the learner is reading through content that is unrelated to the circulatory system and so, for instance, the researcher examining what content the learner is reviewing?) to guide their instructional decision making? Eye movements may also indicate the how extensively information is processed during learning (Catrysse et al., 2018). For instance, when participants reported both deep and surface-level information processing, they tended to fixate longer and revisit content more often than participants who reported only surface-level processing (Catrysse et al., 2018); but see (Winne, 2018b, 2020) for a critique of the "depth" construct).

Other studies have used think-aloud protocols for data mining to seek emic descriptions of information processing and reasoning (Greene et al., 2018). Muldner et al. (2010) drew inferences from concurrent verbalizations representing self-explaining, describing connections between problems or examples, and other key cognitive processes (e.g., summarizing content) in Physics during learning with an intelligent tutoring system. Similar analytic approaches were used to understand clinicians' diagnostic reasoning (Kassab and Hussain, 2010). Si et al. (2019) used a rubric to quantify the quality of their participants' reasoning about a diagnosis. Quality of reasoning was positively related to clinical-reasoning skills and accurate diagnoses. These findings indicate that think-aloud methods can quantify how and when researchers analyze and reason about learners' multimodal data (Si et al., 2019). Multimodal data about researchers' engagements with learners' multimodal data can inform where researchers' monitoring and analyzing behaviors about deciding if, when, and how to scaffold the learners' SRL. Negative feedback loops built into SPARC (see Figure 3) offer pathways for efficiently examining researchers' understanding of learners' SRL, and how the researchers' biases related to their beliefs about SRL and the effectiveness of their instructional decision-making. Overall, the studies reviewed here illustrate compounding of value by coordinating think-aloud protocols, eye-gaze data, and online behavioral traces to capture implicit processes such as analyzing and reasoning. Therefore, the SPARC system should be engineered to capture and mine patterns within researchers' concurrent verbalizations, eye movements, and clickstream data to mark with what, when, how, and how long researchers reason and analyze learners' multimodal data as they forge inferences about learners' SRL. However, data streams sampling researchers' activation of monitoring processes and marking instances of analyzing and reasoning merely set a stage for inquiring whether researchers

understand how learners' multimodal data represent SRL. Simply tracking researchers' metacognition is insufficient to guide instructional decision-making that optimizes scaffolding learners' SRL. Researchers' understanding is also necessary.

2.3. Understanding real-time multimodal data

People acquire conceptual knowledge by coordinating schema and semantic networks to encode conceptual and propositional knowledge (Anderson, 2000). Therefore, researchers' understanding of learners' SRL represented by multimodal data depends on access to valid schemas and slots within them, and a well-formed structure of networked information about learners' SRL. For example, (Mudrick et al., 2019) results indicate a learner's eye fixations oscillating between text and a diagram (i.e., saccades) should fill a slot in a schema for metacognitive monitoring within a schema describing motivation to build comprehension by, for this slot, resolving confusion. SPARC should detect whether an instructional decision-maker activates and instantiates schemas like this. Then, merging that information and other data about learners and researchers into negative feedback within a pipeline architecture, the system can iteratively scaffold researchers toward successively improved decisions about interventions that optimize learners' performance and self-regulation.

2.3.1. Data channels capturing understanding

Traditionally, comprehension has been assessed using aggregated total gain scores drawn from selected-response, paper, and pencil tests before and after domain-specific instruction (Makransky et al., 2019). However, process-oriented and performance-based methods using multimodal data offer promising alternatives (James et al., 2016). Liu et al. (2019) sampled multiple data streams over a learning session using video and audio recordings, physiological sensors, eye tracking, and online behavioral traces. Their model formed from multimodal process-based data more strongly predicted learners' understanding than models based on a single modality of data such as online behavioral traces (Liu et al., 2019). Similarly, Makransky et al. (2019) amalgamated multimodal data across online behavioral traces, eye tracking, electrophysiological signals, and heart rate to build models predicting variance in learners' understanding of information taught during a learning session. A unimodal model using just online behavioral traces explained 57% of the variance ($p < 0.05$) in learners' understanding. A model incorporating multiple data streams explained 75% of the variance ($p < 0.05$) in learners' understanding (Makransky et al., 2019). Thus, in order for SPARC to capture researchers' understanding of learners' SRL using multimodal data, the system would need to sample various data channels to learn data indices that indicate if, when, and how the researcher is understanding the learners' SRL. Depending on the researcher's multimodal data and their accuracy in understanding a learner's SRL, over time SPARC would learn each researcher's understanding of learners' data so that the system may make accurate inferences about when to scaffold researchers to optimize their understanding of learners' multimodal data. We suggest that in order to capture researchers' understanding of information, SPARC could be built to automatically capture understanding based on using eye-gaze behaviors, concurrent verbalizations, mouse-clicks (e.g., what is the researcher and/or learner attending to, and is the action related to the objective of the session—for instance, is the researcher attending to data channels or modalities signaling an SRL process that needs scaffolding, such that of a learner attending to text and diagrams that are irrelevant to the learning objective in which they are studying. Does the researcher monitor these data channels to guide their instructional decision-making, and if so, how does the instructional decision impact the learners' subsequent SRL during the session? Using SPARC to sample and model the multimodal data generated by both researchers and learners could help answer these questions and advance research on the science of learning.

3. Results

3.1. Theoretically- and empirically-based system guidelines for SPARC

SPARC should be engineered to offload tasks that overload researchers' attention and working memory to key analytics describing learners' SRL and integration of those analytics in forming productive instructional decisions. We describe here how SPARC can address these critical needs. Some of SPARC's functionality can be adapted from existing multimodal analytical tools such as SSI (Wagner et al., 2013; Noroozi et al., 2019). SSI's machine-learning pipeline automatically aligns, processes, and filters multimodal data as they are generated in real-time. SPARC will incorporate this functionality and augment it according to theoretical frameworks and empirical findings mined from learning science. Specifically, SPARC's pipeline will be calibrated to weight data channels (e.g., eye gaze, concurrent verbalizations), modalities (e.g., fixation vs. saccades in eye gaze, frequency vs. duration of fixations, etc.), and combinations of data to reflect meaningful and critical learning and SRL processes. For example, facial expressions and physiological sensor data would be assigned greater weight in modeling affect and affective state change, such as frustration, while screen recordings, concurrent verbalizations, and eye-movement data would be assigned greater weights to model cognitive strategy use).

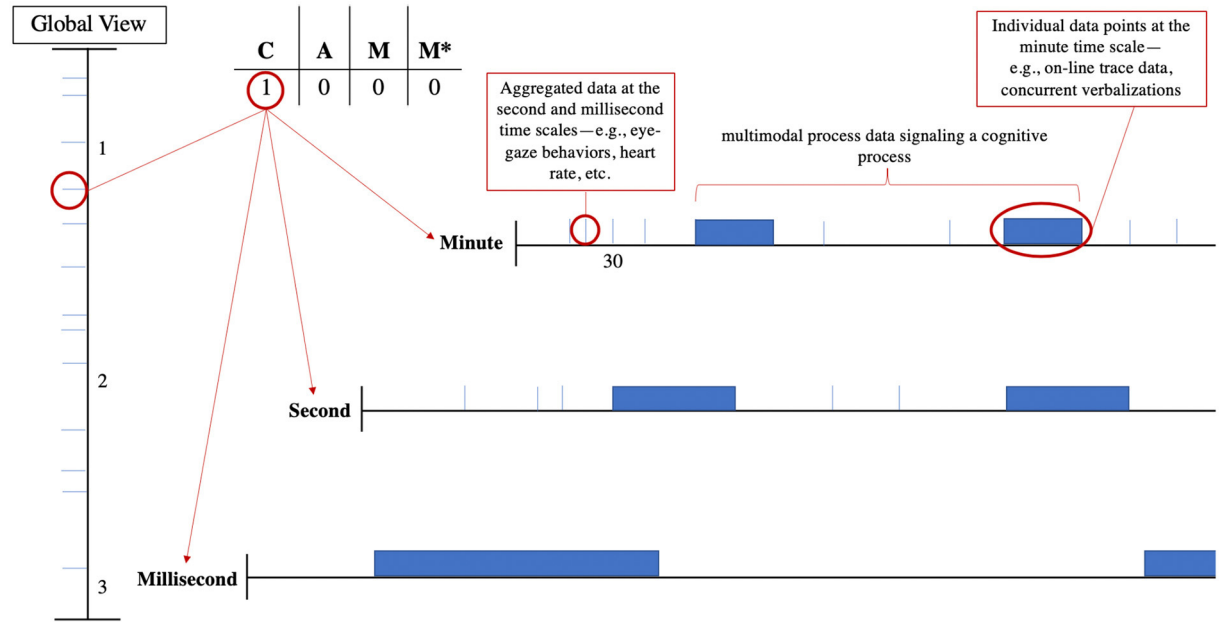


FIGURE 5
Scaling from a global view of SRL constructs at the hour temporal scale to the minute, second, and millisecond temporal scale of signals. C, cognitive; A, affective; M, meta-cognitive; M*, motivation.

Moreover, the SPARC system will use automated-recognition routines to detect and classify learners' and researchers' SRL activities separately while analyzing data from the activities concurrently to guide the scaffolding of the researcher and assessment of how the instructional prescriptions of the researcher are impacting the learners' SRL and performance. For example, when capturing conditions marked by the COPES model (Winne, 2018a), eye-gaze, and think-aloud data may best indicate conditions learners perceive about a learning task and the learning environment. SPARC's algorithms would assign these data greater weight compared to clickstream and physiological sensor data to represent conditions from the learner's perspective. It is important to note algorithms should reflect a full scan of conditions regarding signals about conditions present and conditions absent (Winne, 2019). Theory plays a key role here because it is the source for considering potential roles for a construct that has zero value in the data vector. Temporal dimensions (see Figure 5) are a critical feature in SPARC's approach to modeling a learner's multimodal data considering that different learning processes may unfold across varying time scales. Data within and across channels collected over time helps to ensure adequate sampling (e.g., how long does an affective state last?) and multiple contextual cues (e.g., what did the learner do before and after onset of an affective state?). This wider context enhances interpretability beyond single-channel, single time-point data. For example, a 250 Hz eye tracker supplying 250 data points per second may be insufficient to infer learner processing in the one-second sample. Other data, e.g., sequence of previewing headings, reading, and re-reading indicating multiple metacognitive judgments augmented by screen recording and concurrent verbalizations across several minutes provide a more complete structure for a researcher to draw inferences about the learner's engagement in a task (Mudrick et al., 2019; Taub and Azevedo, 2019).

As such, SPARC features will allow a researcher to scale up—i.e., scale upsampling rates to a uniform temporal scale such as from milliseconds to seconds, or seconds to minutes, or down—i.e., scale downsampling rates to a uniform temporal scale such as from seconds to milliseconds, to pinpoint how, when, why, and what learning processes were occurring (Figure 5). When researchers scale up or down, it also captured critical information revealing how the researcher is selecting, monitoring, analyzing, and understanding learners' multimodal data representing operations in the COPES model and modulations of operations that represent SRL. The opportunity for the researcher to explore the learner's temporal learning progression is a critical feature that researchers need to guide their instructional decision making related to adaptive scaffolding and feedback to the learners (Kinnebrew et al., 2014, 2015; Basu et al., 2017). For SPARC to continuously capture data and update its models of learners and researchers, it should apply predictive models to track the learners' trajectories and project future learning events prompted by researcher intervention. For instance, if the researcher gave learners feedback and redirected a learner to another section content more relevant

to learning objectives, SPARC should forecast the probability of learner uptake of that recommendation and patterns of multimodal data that confirm uptake. Iterating across learning sessions, this allows SPARC to dynamically converge models to more accurately predict both behavior by the learner and the researcher.

4. Discussion

Emerging research on SRL sets the stage for using temporally sequenced multimodal data to examine the dynamics of multiple processes and interventions to adapt those processes in emerging technologies. Using large volumes of multimodal data to analyze and interpret learners' SRL processes in near real-time is theoretically and algorithmically challenging (Cloude et al., 2020; Emerson et al., 2020). We crafted a theoretical, conceptual, and empirically grounded framework for designing a system that guides researchers and instructors in analyzing and understanding the complex nature of SRL. A novel aspect of the SPARC system is modeling all players in instruction—learner and instructional decision maker—to dynamically upgrade capabilities to enhance learning, SRL, and the empirical foundations for understanding those processes. Further, including insights gained from data collected on researchers and SRL experts could potentially contribute to enhancing our understanding of how to automatically build detectors of SRL processes on both instructors and learners. Emerging research using multimodal data shows promise in approaching this goal, but this research stream has not yet tackled major challenges facing interdisciplinary and international researchers and instructors in monitoring, analyzing, and understanding learners' SRL multimodal data based on what, where, when, how, and with what learners self-regulate to understand content. In particular, the SPARC system we outline defines and sets a framework for addressing a new and fundamental question. How do researchers and instructors monitor, analyze, and understand learners' and groups of learners' multimodal data; and, how can data about those processes be merged with data about learners to bootstrap the full system involving learners, instructional decision makers, and interventions? The SPARC system we suggest takes the first steps toward addressing major conceptual, theoretical, methodological, and analytical issues associated with using real-time multimodal data (Winne, 2022).

4.1. Implications and future directions

Implications of this research are threefold. First, leveraging insights gained from researchers' and SRL experts' multimodal data based on their understanding of both (1) instructors' and (2) learners' multimodal data could be used to build *valid* algorithms for SRL detection. For example, the SRL expert could potentially tag whether the instructor identified the learners' misuse of SRL while viewing materials? If the instructor did identify this behavior, did they intervene accordingly based on their own SRL and understanding of the learner's processes to make an informed instructional decision? Utilizing the information that the SRL expert or researcher referenced could be used to build SRL detectors. Training AI on how researchers, instructors, and learners monitor, infer, and understand information across multiple data sources has the potential to build valid algorithms that are empirically and theoretically based derived. Building valid AI is a current challenge for the field, where most AI is built by experts that have little knowledge about SRL theory. Instead, AI algorithms are data-driven such that the steps are built to maximize the detection of significant findings with the highest accuracy. This approach slows progress on deriving meaningful insights from relationships present in multiple data sources. Through utilizing SPARC, it would ensure that the best algorithms/data channels/modalities/dependent variables are selected based on a combination of the researchers', instructors', and learners' information as a whole. Furthermore, this would also spark researchers, instructors, and learners to think critically about what the algorithm should be doing to facilitate understanding of SRL for supporting informed instructional decision-making. This research may highlight areas for teaching training, such as integrating data science and visualizations courses in the curriculum since data are being increasingly used in the classroom to enhance the quality of education. Was this monitoring or behavior? If not, why did the algorithm fire to suggest it was so? It could provide a world of information about where the researcher is doing quality control on the algorithm to assess if they are working properly in all contexts. This could generate a library of open-source algorithms/production rules for a range of contexts, domains, users, countries, theories, and many others. Another important area is leveraging SPARC to reveal user biases. For example, is an instructor focusing on specific data sources or all data sources? Is the instructor supporting all learners in the same way? SPARC would allow us to compare and contrast where users could be biased toward certain data channels relative to others, and potential shed light on these behaviors to mitigate bias and draw awareness to our perspectives when we are not using SPARC, thus potentially enhancing our objectivity as scientists and instructors.

One area of future research that could advance this work is moving away from solely relying on a linear paradigm to define SRL such as linear regression. It is imperative that we utilize sophisticated statistical techniques to model the complexity and dynamics that emerge within multimodal data across varying system levels such as multimodal data collected from the researcher or instructor in their understanding of analytics presented back to the user. An interdisciplinary approach toward data processing and analysis may provide the analytical

tools needed to exploit meaningful relationships and insights within the data. Specifically, we need to go beyond information-processing theory which assumes that self-regulated learning results from linear sequences of learning processes as assumed in linear models. We challenge research to make a paradigmatic shift toward dynamic systems thinking (Van Gelder and Port, 1995) to investigate researchers', instructors', and learners' SRL processes as self-organizing, dynamically emergent, and non-linear phenomena. Leaning on nonlinear dynamical analyses to study SRL is starting to gain momentum (Dever et al., 2022; Li et al., 2022). This interdisciplinary approach would allow us to study SRL across multiple levels and nonlinear dynamical analyses offer more flexibility in utilizing multiple data sources that do not need to adhere to rigid assumptions of normality, equal variation, and independence of observation. Finally, SPARC offers implications for building AI-enabled adaptive learning systems that repurpose information back to instructors, learners, and research to augment both teaching and learning Kabudi et al. (2021). As outlined by Kabudi et al. (2021), AI-enabled adaptive learning systems could detect and select the appropriate learning intervention using evidence from SRL experts, researchers, instructors, and learners. data collected during learning activities should not only be predictive analytics but rather leverages data in various ways depending on the (a) user and (b) objective of the session. Specifically, analytics fed back to users such as instructors should include both prescriptive and descriptive analytics. (go into prescriptive, descriptive, and predictive; Kabudi et al., 2021) as these all hold different implications for teaching and learning.

5. Conclusions

Much work remains to realize SPARC. New research is needed to widen and deepen understandings of (1) how to map researchers' and learners' multimodal data onto COPES constructs, (2) how differences in these mappings suggest interventions and the degree to which this may vary by country, and (3) how a pipeline architecture should be designed to iterate over these results to optimize both learner's SRL, instructor's SRL, and instructional decision making. Our next steps in building and implementing a prototype system like SPARC are to begin collecting real-time data about how researchers examine and use learners' multimodal data in specific learning and problem-solving scenarios, e.g., learning with serious games, intelligent tutoring systems, and virtual reality. This requires recruiting a number of leading experts within the field of self-regulated learning across a range of emerging technologies, but also a range of SRL theories including socially-shared self-regulated learning and co-regulation. Each scenario presented to participants (i.e., researchers/experts) will encompass gathering their multimodal data while they review learners' multimodal data and provide annotations that classify various SRL processes and strategies. This study will allow further understanding of how researchers monitor, analyze, and make inferences about SRL using multimodal data. Results will guide system architecture and design for a theoretically- and empirically-based system that supports researchers and instructors in monitoring, analyzing, and understanding learners' multimodal data to make effective instructional decisions and foster self-regulation (Hwang et al., 2020). Through training AI using the multiple data channels collected from leading researchers and experts within the field of self-regulated learning, in conjunction with instructors and learners, it opens opportunities to build valid and reliable AI that goes beyond data-driven techniques to determine when theoretically-relevant constructs emerge across a range of data channels and modalities.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

Ethical review and approval was not required for the current study in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

EC conceptualized, developed, synthesized, and constructed the manuscript. RA assisted in conceptualizing and developing the SPARC vision. PW provided extensive edits and revisions to the manuscript and assisted in conceptualizing the SPARC vision. GB provided edits and revisions to the manuscript. EJ provided edits and revisions to the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was partially supported by a grant from the National Science Foundation (DRL#1916417) awarded to RV.

Acknowledgments

The authors would like to thank team members of the SMART Lab at the University

of Central Florida for their assistance and contributions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Ann. N. Y. Acad. Sci.* 1369, 24–39. doi: 10.1111/nyas.12957

Anderson, B. A., and Yantis, S. (2013). Persistence of value-driven attentional capture. *J. Exp. Psychol.* 39, 6. doi: 10.1037/a0030860

Anderson, J. R. (2000). *Cognitive Psychology and Its Implications*. New York, NY: Worth Publishers.

Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., et al. (2022). Lessons learned and future directions of metatutor: leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Front. Psychol.* 13, 813632. doi: 10.3389/fpsyg.2022.813632

Azevedo, R., and Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: issues and challenges. *Comput. Hum. Behav.* 96, 207–210. doi: 10.1016/j.chb.2019.03.025

Azevedo, R., Mudrick, N. V., Taub, M., and Bradbury, A. E. (2019). "Self-regulation in computer-assisted learning systems," in *The Cambridge Handbook of Cognition and Education*, eds J. Dunlosky and K. A. Rawson (Cambridge, UK: Cambridge University Press), 587–618. doi: 10.1017/9781108235631.024

Azevedo, R., Taub, M., and Mudrick, N. V. (2018). "Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies," in *Handbook on Self-Regulation of Learning and Performance*, eds D. H. Schunk and J. A. Greene (New York, NY: Routledge). doi: 10.4324/9781315697048-17

Basu, S., Biswas, G., and Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Model. User Adapt. Interact.* 27, 5–53. doi: 10.1007/s11257-017-9187-0

Bernacki, M. L. (2018). "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*, eds D. H. Schunk and J. A. Greene (New York, NY: Routledge). doi: 10.4324/9781315697048-24

Biswas, G., Segedy, J. R., and Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty's brain. *Int. J. Artif. Intell. Educ.* 26, 350–364. doi: 10.1007/s40593-015-0057-9

Blanchette, I., and Richards, A. (2010). The influence of affect on higher level cognition: a review of research on interpretation, judgement, decision making and reasoning. *Cogn. Emot.* 24, 561–595. doi: 10.1080/02699930903132496

Blascheck, T., Beck, F., Baltes, S., Ertl, T., and Weiskopf, D. (2016). "Visual analysis and coding of data-rich user behavior," in *Proceedings of the 2016 IEEE Conference on Visual Analytics Science and Technology, VAST '16* (New York, NY), 141–150. doi: 10.1109/VAST.2016.7883520

Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Lesterhuis, M., and Van den Bossche, P. (2018). How are learning strategies reflected in the eyes? Combining results from self-reports and eye-tracking. *Br. J. Educ. Psychol.* 88, 118–137. doi: 10.1111/bjep.12181

Chango, W., Cerezo, R., Sanchez-Santillan, M., Azevedo, R., and Romero, C. (2021). Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources. *J. Comput. High. Educ.* 33, 614–634. doi: 10.1007/s12528-021-09298-8

Claypoole, V. L., Dever, D. A., Denues, K. L., and Szalma, J. L. (2019). The effects of event rate on a cognitive vigilance task. *Hum. Factors* 61, 440–450. doi: 10.1177/0018720818790840

Cloude, E. B., Dever, D. A., Wiedbusch, M. D., and Azevedo, R. (2020). Quantifying scientific thinking using multichannel data with crystal island: implications for individualized game-learning analytics. *Front. Educ.* 5, 217. doi: 10.3389/feduc.2020.572546

Cloude, E. B., Wortha, F., Dever, D. A., and Azevedo, R. (2021a). "Negative emotional dynamics shape cognition and performance with metatutor: toward building affect-aware systems," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (New York, NY), 1–8. doi: 10.1109/ACII52823.2021.9597462

Cloude, E. B., Wortha, F., Wiedbusch, M. D., and Azevedo, R. (2021b). "Goals matter: changes in metacognitive judgments and their relation to motivation and learning with an intelligent tutoring system," in *International Conference on Human-Computer Interaction* (Cham: Springer), 224–238. doi: 10.1007/978-3-030-77889-7_15

Desimone, R. (1996). "Neural mechanisms for visual memory and their role in attention," in *Proceedings of the National Academy of Sciences of the United States of America, Vol. 93 of PANAS '96* (Washington, DC: National Academy of Sciences), 13494–13499. doi: 10.1073/pnas.93.24.13494

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Dever, D. A., Amon, M. J., Vrzáková, H., Wiedbusch, M. D., Cloude, E. B., and Azevedo, R. (2022). Capturing sequences of learners' self-regulatory interactions with instructional material during game-based learning using auto-recurrence quantification analysis. *Front. Psychol.* 13, 813677. doi: 10.3389/fpsyg.2022.813677

Dindar, M., Järvelä, S., Nguyen, A., Haataja, E., and Çini, A. (2022). Detecting shared physiological arousal events in collaborative problem solving. *Contemp. Educ. Psychol.* 69, 102050. doi: 10.1016/j.cedpsych.2022.102050

Duncan, J., and Nimmo-Smith, I. (1996). Objects and attributes in divided attention: Surface and boundary systems. *Percept. Psychophys.* 58, 1076–1084.

Emerson, A., Cloude, E. B., Azevedo, R., and Lester, J. (2020). Multimodal learning analytics for game-based learning. *Br. J. Educ. Technol.* 51, 1505–1526. doi: 10.1111/bjet.12992

Franco-Watkins, A. M., Mattson, R. E., and Jackson, M. D. (2016). Now or later? Attentional processing and intertemporal choice. *J. Behav. Decis. Making* 29, 206–217. doi: 10.1002/bdm.1895

Greene, J. A., and Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educ. Psychol.* 45, 203–209. doi: 10.1080/00461520.2010.515935

Greene, J. A., Deekens, V. M., Copeland, D. Z., and Yu, S. (2018). "Capturing and modeling self-regulated learning using think-aloud protocols," in *Handbook of Self-Regulation of Learning and Performance*, eds D. H. Schunk and J. A. Greene (New York, NY: Routledge). doi: 10.4324/9781315697048-21

Greenlee, E. T., DeLucia, P. R., and Newton, D. C. (2019). Driver vigilance in automated vehicles: effects of demands on hazard detection performance. *Hum. Factors* 61, 474–487. doi: 10.1177/0018720818802095

Hadwin, A. F. (2021). Commentary and future directions: What can multi-modal data reveal about temporal and adaptive processes in self-regulated learning?. *Learn. Instruct.* 72, 101287. doi: 10.1016/j.learninstruc.2019.101287

Hancock, P. A. (2013). In search of vigilance: the problem of iatrogenically created psychological phenomena. *Am. Psychol.* 68, 97. doi: 10.1037/a0030214

Hawkins, W. J., Heffernan, N. T., and Baker, R. S. (2014). "Learning Bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities," in *International Conference on Intelligent Tutoring Systems* (Cham: Springer), 150–155. doi: 10.1007/978-3-319-07221-0_18

Hwang, G.-J., Xie, H., Wah, B. W., and Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Comput. Educ.* 1, 100001. doi: 10.1016/j.caeai.2020.100001

Ifenthaler, D., and Schumacher, C. (2019). "Releasing personal information within learning analytics systems," in *Learning Technologies for Transforming Large-Scale Teaching, Learning, and Assessment*, eds D. Sampson, J. Spector, D. Ifenthaler, P. Isaías, and S. Sergis (Cham: Springer), 3–18. doi: 10.1007/978-3-030-15130-0_1

Ifenthaler, D., and Tracey, M. W. (2016). Exploring the relationship of ethics and privacy in learning analytics and design: implications for the field of educational technology. *Educ. Technol. Res. Dev*. 64, 877–880. doi: 10.1007/s11423-016-9480-3

James, P., Antonova, L., Martel, M., and Barkun, A. N. (2016). Measures of trainee performance in advanced endoscopy: a systematic review: 342. *Am. J. Gastroenterol*. 111, S159. doi: 10.14309/00000434-201610001-00342

Jang, E. E., Lajoie, S. P., Wagner, M., Xu, Z., Poitras, E., and Naismith, L. (2017). Person-oriented approaches to profiling learners in technology-rich learning environments for ecological learner modeling. *J. Educ. Comput. Res*. 55, 552–597. doi: 10.1177/0735633116678995

Johnson-Glenberg, M. C. (2018). Immersive VR and education: embodied design principles that include gesture and hand controls. *Front. Robot. AI* 5, 81. doi: 10.3389/frobt.2018.00081

Kabudi, T., Pappas, I., and Olsen, D. H. (2021). Ai-enabled adaptive learning systems: a systematic mapping of the literature. *Comput. Educ*. 2, 100017. doi: 10.1016/j.caeai.2021.100017

Kassab, S. E., and Hussain, S. (2010). Concept mapping assessment in a problem-based medical curriculum. *Med. Teach*. 32, 926–931. doi: 10.3109/0142159X.2010.497824

Kinnebrew, J. S., Mack, D. L., Biswas, G., and Chang, C.-K. (2014). "A differential approach for identifying important student learning behavior patterns with evolving usage over time," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD '14*, eds W. C. Peng, et al. (Cham: Springer), 281–292. doi: 10.1007/978-3-319-13186-3_27

Kinnebrew, J. S., Segedy, J. R., and Biswas, G. (2015). Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. *IEEE Trans. Learn. Technol*. 10, 140–153. doi: 10.1109/TLT.2015.2513387

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/7688.001.0001

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., and Gube, M. (2021). Examining the interplay of affect and self-regulation in the context of clinical reasoning. *Learn. Instruct*. 72, 101219. doi: 10.1016/j.learninstruc.2019.101219

Lane, H. C., and D'Mello, S. K. (2019). "Uses of physiological monitoring in intelligent learning environments: a review of research, evidence, and technologies," in *Mind, Brain and Technology: Learning in the Age of Emerging Technologies*, eds T. D. Parsons, L. Lin, and D. Cockerham (Cham: Springer International Publishing), 67–86. doi: 10.1007/978-3-030-02631-8_5

Li, S., Zheng, J., Huang, X., and Xie, C. (2022). Self-regulated learning as a complex dynamical system: examining students' stem learning in a simulation environment. *Learn. Individ. Differ*. 95, 102144. doi: 10.1016/j.lindif.2022.102144

Liu, S., Ji, H., and Wang, M. C. (2019). Nonpooling convolutional neural network forecasting for seasonal time series with trends. *IEEE Trans. Neural Netw. Learn. Syst*. 95. doi: 10.1109/TNNLS.2019.2934110

Makransky, G., Terkildsen, T. S., and Mayer, R. E. (2019). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learn. Instruct*. 61, 23–34. doi: 10.1016/j.learninstruc.2018.12.001

Malmberg, J., Järvelä, S., and Järvenoja, H. (2017). Capturing temporal and sequential patterns of self-, co-, and socially shared regulation in the context of collaborative learning. *Contemp. Educ. Psychol*. 49, 160–174. doi: 10.1016/j.cedpsych.2017.01.009

Mayer, R. E. (2019). Thirty years of research on online learning. *Appl. Cogn. Psychol*. 33, 152–159. doi: 10.1002/acp.3482

Messi, D. and Adrario, E. (2021). "Mixed reality simulation for medical training: how it affects learners; cognitive state," in *Advances in Simulation and Digital Human Modeling: Proceedings of the AHFE 2021 Virtual Conferences on Human Factors and Simulation, and Digital Human Modeling and Applied Optimization, July 25-29, 2021, USA, Vol. 264* (Cham: Springer Nature), 339. doi: 10.1007/978-3-030-79763-8_41

Molenaar, I. and Järvelä, S. (2014). Sequential and temporal characteristics of self and socially regulated learning. *Metacogn. Learn*. 9, 75–85. doi: 10.1007/s11409-014-9114-2

Mudrick, N. V., Azevedo, R., and Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Comput. Hum. Behav*. 96, 223–234. doi: 10.1016/j.chb.2018.06.028

Muldner, K., Burleson, W., and VanLehn, K. (2010). "yes!: using tutor and sensor data to predict moments of delight during instructional activities," in *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization, UMAP '10*, eds P. De Bra, A. Kobsa, and D. Chin (New York, NY: Springer), 159–170. doi: 10.1007/978-3-642-13470-8_16

Nguyen, P. H., Xu, K., Wheat, A., Wong, B. W., Attfield, S., and Fields, B. (2015). Sensepath: understanding the sensemaking process through analytic provenance. *IEEE Trans. Visual. Comput. Graph*. 22, 41–50. doi: 10.1109/TVCG.2015.2467611

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., and Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Comput. Hum. Behav*. 100, 298–304. doi: 10.1016/j.chb.2018.12.019

Raca, M., and Dillenbourg, P. (2014). "Holistic analysis of the classroom," in *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge, MLA '14* (New York, NY: ACM), 13–20. doi: 10.1145/2666633.2666636

Rajendran, R., Kumar, A., Carter, K. E., Levin, D. T., and Biswas, G. (2018). "Predicting learning by analyzing eye-gaze data of reading behavior," in *Proceedings of the 11th International Conference on Educational Data Mining* (New York, NY).

Reimann, P. (2021). Methodological progress in the study of self-regulated learning enables theory advancement. *Learn. Instruct*. 72, 101269. doi: 10.1016/j.learninstruc.2019.101269

Renninger, K. A., and Hidi, S. E. (2019). "Interest development and learning," in *The Cambridge Handbook of Motivation and Learning*, eds K. A. Renninger and S. E. Hidi (Cambridge, UK: Cambridge University Press). doi: 10.1017/9781316823279.013

Salomon, G., and Perkins, D. N. (1989). Rocky roads to transfer: rethinking mechanism of a neglected phenomenon. *Educ. Psychol*. 24, 113–142. doi: 10.1207/s15326985ep2402_1

Scheiter, K., and Eitel, A. (2017). "The use of eye tracking as a research and instructional tool in multimedia learning," in *Eye-Tracking Technology Applications in Educational Research* (Hershey, PA: IGI Global), 143–164. doi: 10.4018/978-1-5225-1005-5.ch008

Schunk, D. H., and Greene, J. A. (2018). *Handbook of Self-Regulation of Learning and Performance*. New York, NY: Routledge. doi: 10.4324/9781315697048

Segedy, J. R., Kinnebrew, J. S., and Biswas, G. (2015). Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *J. Learn. Analyt*. 2, 13–48. doi: 10.18608/jla.2015.21.3

Si, J., Kong, H.-H., and Lee, S.-H. (2019). Developing clinical reasoning skills through argumentation with the concept map method in medical problem-based learning. *Interdiscip. J. Probl. Based Learn*. 13, 1–16. doi: 10.7771/1541-5015.1776

Spires, H. A., Rowe, J. P., Mott, B. W., and Lester, J. C. (2011). Problem solving and game-based learning: effects of middle grade students' hypothesis testing strategies on learning outcomes. *J. Educ. Comput. Res*. 44, 453–472. doi: 10.2190/EC.44.4.e

Taub, M., and Azevedo, R. (2019). How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive SRL processes during learning with an intelligent tutoring system? *Int. J. Artif. Intell. Educ*. 29, 1–28. doi: 10.1007/s40593-018-0165-4

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., and Price, M. J. (2021). How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?. *Learn. Instruct*. 72, 101200. doi: 10.1016/j.learninstruc.2019.04.001

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., and Lester, J. (2016). "Using multi-level modeling with eye-tracking data to predict metacognitive monitoring and self-regulated learning with crystal island," in *Proceedings of the International Conference on Intelligent Tutoring Systems, ITS '16* (New York, NY: Springer), 240–246. doi: 10.1007/978-3-319-395 83-8_24

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., and Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with crystal island. *Comput. Hum. Behav*. 76, 641–655. doi: 10.1016/j.chb.2017.01.038

UNESCO (2017). *More Than One-Half of Children and Adolescents Are Not Learning Worldwide. UIS Fact Sheet No. 46*. UNESCO.

Van Gelder, T., and Port, R. F. (eds.). (1995). "It's about time: an overview of the dynamical approach to cognition," in *Mind as Motion: Explorations in the Dynamics of Cognition, Vol. 1* (Cambridge, MA: The MIT Press), 1–44.

van Zoest, W., Van der Stigchel, S., and Donk, M. (2017). Conditional control in visual selection. *Attent. Percept. Psychophys*. 79, 1555–1572. doi: 10.3758/s13414-017-1352-3

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., and André, E. (2013). "The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM international conference on Multimedia*, 831–834. doi: 10.1145/2502081.2502223

Wiedbusch, M., Dever, D., Wortha, F., Cloude, E. B., and Azevedo, R. (2021). "Revealing data feature differences between system-and learner-initiated self-regulated learning processes within hypermedia," in *International Conference on Human-Computer Interaction* (Cham: Springer), 481–495. doi: 10.1007/978-3-030-77857-6_34

Winne, P. (2018). Paradigmatic issues in state-of-the-art research using process data. *Frontl. Learn. Res*. 6, 250–258. doi: 10.14786/flr.v6i3.551

Winne, P. H. (2018a). "Cognition and metacognition within self-regulated learning," in *Handbook of Self-Regulation of Learning and Performance*, eds D. H. Schunk and J. A. Greene (New York, NY: Routledge), 52–64. doi: 10.4324/9781315697048-3

Winne, P. H. (2018b). Theorizing and researching levels of processing in self-regulated learning. *Br. J. Educ. Psychol*. 88, 9–20. doi: 10.1111/bjep.12173

Winne, P. H. (2019). Paradigmatic dimensions of instrumentation and analytic methods in research on self-regulated learning. *Comput. Hum. Behav*. 96, 285–289. doi: 10.1016/j.chb.2019.03.026

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Comput. Hum. Behav*. 112. doi: 10.1016/j.chb.2020.106457

Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: how trace data and learning analytics help develop skills for self-regulated learning. *Metacogn. Learn.* 1–19. doi: 10.1007/s11409-022-09305-y

Winne, P. H., and Hadwin, A. F. (1998). "Studying as self-regulated learning," in *Metacognition in Educational Theory and Practice,* eds D. J. Hacker, J. Dunlosky, and A. C. Graesser (Mahwah, NJ: Lawrence Erlbaum Associates), 277–304.

Winne, P. H., Teng, K., Chang, D., Lin, M. P.-C., Marzouk, Z., Nesbit, J. C., et al. (2019). nStudy: software for learning analytics about processes for self-regulated learning. *J. Learn. Anal.* 6, 95–106. doi: 10.18608/jla.2019.62.7

frontiers | Frontiers in Psychology

TYPE Perspective
PUBLISHED 25 August 2022
DOI 10.3389/fpsyg.2022.956798

OPEN ACCESS

EDITED BY
Iza Marfisi-Schottman,
EA4023 Laboratoire d'Informatique de l'Université du Mans (LIUM), France

REVIEWED BY
Ranilson Oscar Araújo Paiva,
Federal University of Alagoas, Brazil
Farzan Shenavarmasouleh,
University of Georgia, United States

*CORRESPONDENCE
Manuel Ninaus
manuel.ninaus@uni-graz.at

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

RECEIVED 30 May 2022
ACCEPTED 08 August 2022
PUBLISHED 25 August 2022

CITATION
Ninaus M and Sailer M (2022) Closing the loop – The human role in artificial intelligence for education. *Front. Psychol.* 13:956798. doi: 10.3389/fpsyg.2022.956798

Closing the loop – The human role in artificial intelligence for education

Manuel Ninaus[1,2*†] and Michael Sailer[3†]

[1]Institute of Psychology, University of Graz, Graz, Austria, [2]LEAD Graduate School and Research Network, University of Tübingen, Tübingen, Germany, [3]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Recent advancements in artificial intelligence make its use in education more likely. In fact, existing learning systems already utilize it for supporting students' learning or teachers' judgments. In this perspective article, we want to elaborate on the role of humans in making decisions in the design and implementation process of artificial intelligence in education. Therefore, we propose that an artificial intelligence-supported system in education can be considered a closed-loop system, which includes the steps of (i) data recording, (ii) pattern detection, and (iii) adaptivity. Besides the design process, we also consider the crucial role of the users in terms of decisions in educational contexts: While some implementations of artificial intelligence might make decisions on their own, we specifically highlight the high potential of striving for hybrid solutions in which different users, namely learners or teachers, are provided with information from artificial intelligence transparently for their own decisions. In light of the non-perfect accuracy of decisions of both artificial intelligence-based systems and users, we argue for balancing the process of human- and AI-driven decisions and mutual monitoring of these decisions. Accordingly, the decision-making process can be improved by taking both sides into account. Further, we emphasize the importance of contextualizing decisions. Potential erroneous decisions by either machines or humans can have very different consequences. In conclusion, humans have a crucial role at many stages in the process of designing and using artificial intelligence for education.

KEYWORDS

technology enhanced learning, artificial intelligence (AI), machine learning (ML), adaptivity, digital technologies, education

Introduction

Imagine participating in an online course hosted on an automated AI-supported learning management system (LMS). After you have completed the latest chapter, the LMS points out that you failed to understand a specific issue in the learning material. Consequently, the system automatically repeats the latest course section you

had already studied. Critically, the judgment of the system is wrong. Such a situation might demotivate you to continue with the course, or you might have lost your trust in the system. The AI-supported LMS drew wrong conclusions based on the available data about you and your learning process, recognized an incorrect pattern in your data, and failed to adapt the system to your actual needs.

With this simplified example of a learning situation in digital learning environments, we want to illustrate that AI-based systems typically do not have 100% accuracy in their judgment. This might lead to devastating results on the learners' or the teachers' end. In the current article, we want to emphasize that the accuracy of predictions of AI-based systems depends on several steps that make up such a system and that humans can and should play a critical role as decision-makers along those steps and along the learning process. Specifically, we argue that AI-supported learning systems can be described as a closed-loop system (see **Figure 1**) as we know it from other feedback-rich learning systems such as neurofeedback (e.g., Ninaus et al., 2013), brain-computer interfaces (e.g., Liarokapis et al., 2014; Kober et al., 2018) and learning analytics systems (e.g., Clow, 2012). In particular, we suggest a closed-loop system for AI-supported learning systems, which consists of the following steps: (i) data recording, (ii) pattern detection, (iii) adaptivity. In the following, we will briefly highlight each of those steps with a particular emphasis on the critical role of humans.

Data recording

Today's hardware, network technologies, and data processing methods allow for the recording and processing of highly heterogeneous and multi-modal data (e.g., Di Mitri et al., 2018). Sensors can provide us not only with contextual data such as time, temperature, or location, but also with very personal data. The latter can be divided into behavioral (e.g., "clicks," comments, time spent on a page) and physiological data (e.g., heart rate, electrodermal activity, brain activity). These data are particularly well suited for mapping processes because they can be recorded at a high sampling rate. Accordingly, the data can provide a (more) comprehensive picture of the learning process itself (for a review see Baker et al., 2020).

Nowadays, many people already use physiological sensors to track physical activity (for a review see, e.g., Gal et al., 2018). In contrast, the use of physiological and behavioral data to record and optimize learning activities is still rare in learning contexts, especially related to personalization of learning tasks in real-time. Undoubtedly, this will change in the future, as a growing number of studies show that physiological and behavioral data of learners are valuable for generating user models and fostering learning (for reviews see Mangaroska and Giannakos, 2019; Ninaus and Nebel, 2021). For instance, Li et al. (2020) used behavioral clickstream data from an LMS to predict performance in a course. Appel et al. (2021), on the other hand, used eye movement parameters to predict learners' cognitive load in a game-based simulation. Compared to traditional performance data available after completing a learning task (e.g., scores, grades), continuously recorded physiological and behavioral data can provide deeper insight into cognitive, emotional, and motivational processes.

Even if the pure recording of data is automatic and thus purely machine-based, humans as decision-makers play a crucial role in (i) selecting appropriate sensors and metrics promising for the learning context, (ii) choosing data to be recorded, and (iii) implementing hardware and software architecture to record the data (see Di Mitri et al., 2018). In all of these steps, data handling has to be considered to be sustainable, responsible, and ethical (for a comprehensive discussion see Hakimi et al., 2021). This includes the transparency of data collection, appropriate communication with relevant stakeholders (see Drachsler and Greller, 2016), the use of established theoretically sound approaches for data selection, and the recording of data that indeed has the potential to foster learning. These aspects require expertise from a wide range of disciplines, such as computer science, psychology, and educational science as well as the collaboration between practitioners and researchers.

Pattern detection

The selection of sensors and data to be recorded leads directly to the next step in our closed-loop system. Learning is a complex and dynamic process. Thus, it is unlikely to map and explain such a process using single data points, such as exam grades or a summative score. Accordingly, large amounts of data are necessary to better understand the learning process. However, as human perception and processing capacity cannot monitor numerous data sources simultaneously, interpretation of large amounts of data and metrics is difficult. Therefore, the focus of the next step in the loop is the identification of patterns in data using ML methods. Specifically, establishing a relationship between different parts of data (e.g., interaction duration with certain learning material) and a target variable (e.g., correct response).

For example, Brandl et al. (2021) recorded each click in a simulation for learning to diagnose patients with diseases. They were able to predict correct or incorrect diagnoses by using ML algorithms. The ML algorithm was used to identify activities that had the greatest influence on correct or incorrect diagnoses. In another study, automated facial emotion detection together with ML was used to classify whether individuals engaged in a game-based or a non-game-based mathematics learning task (Ninaus et al., 2019). Even though the prediction was successful, the used ML algorithm did not provide information on which emotions or magnitude thereof were relevant for successful prediction.

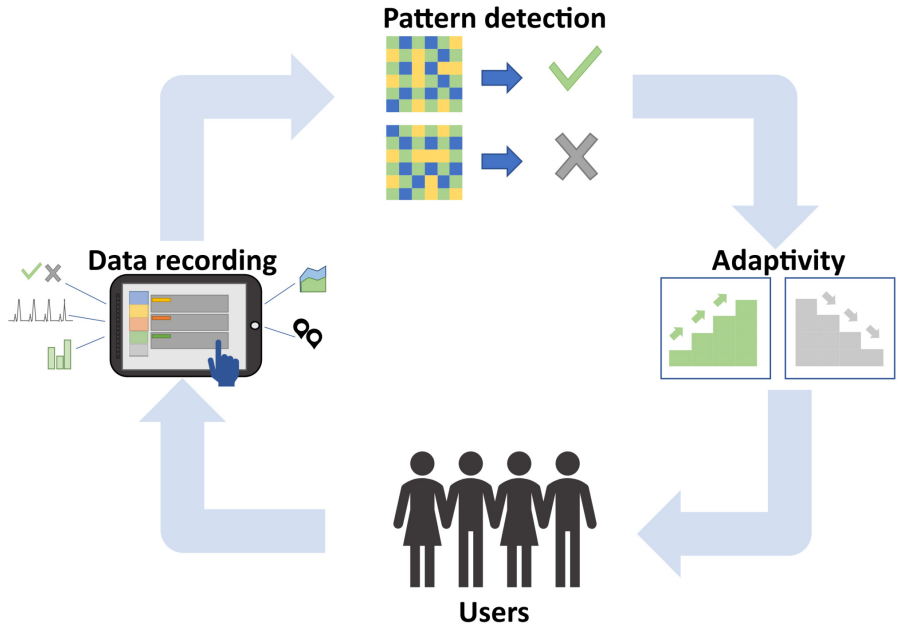


FIGURE 1
Closed-loop system for AI in education, including the steps (i) data recording, (ii) pattern detection, and (iii) adaptivity.

In both of these studies, ML was used to identify patterns in the recorded data. However, their approaches and interpretability of the results differed clearly. This can be partially attributed to the ML algorithm used (Random Forest Model vs. Support Vector Machine). The selection and decision for or against a particular ML algorithm is another key aspect in AI-supported learning systems, which should not only be data-driven but also informed by theory and determined by the overall goal.

Furthermore, differences between supervised and unsupervised ML algorithms should also be considered. The primary goal of supervised ML is to establish a relationship between different parts of the data (e.g., different activities in a simulation) and a target variable (e.g., correct/incorrect response; see Brandl et al., 2021). In unsupervised ML methods, the focus is on exploratory data analysis and clustering of data. Typically, there is no specific outcome variable, such as study success. Instead, one of the aims is to identify subgroups from a set of existing data which can be used for further analysis (e.g., Huijsmans et al., 2020).

However, as mentioned above, learning is a complex and dynamic process. Thus, learning processes might not be simple enough to be represented in a model that humans can always understand (for a comprehensive discussion, see Yarkoni and Westfall, 2017). For instance, ML and AI could be used to predict dropout rates in college or learning success for a course, but the underlying mechanisms might remain hidden from us. Nevertheless, the recent trend toward interpretable ML addresses the criticism of conventional ML of merely providing predictions and emphasizes transparency of the inner workings of ML models to better understand ML-guided decision-making (for a deeper methodological discussion, see Hilbert et al., 2021). This is especially relevant when studying learning processes, as it is crucial to find out which individual variables or aspects of an intervention positively or negatively influence learning success. This information can inform and influence the adaptation of a digital learning environment.

Adaptivity

The next step in our proposed loop concerns the question of how the automatically detected pattern can be used in a learning environment to foster learning. One option is to directly provide detected information to different stakeholders involved in the process: learners and teachers. Learners can receive information about the detected sequences or patterns as feedback on the current performance. This information might be further processed in digital learning environments and provide learners with suggestions on how to adapt to certain problems that might have occurred in their learning process (see Plass and Pawar, 2020).

Similarly, teachers can also receive information about detected sequences and patterns of the learners' learning process. This can help them improve their judgments based on the information received and eventually initiate support. One way is the use of teacher dashboards, which provide teachers with elaborated information about students' learning processes.

Further, teacher dashboards can automatically suggest support measures for specific learners (see Wiedbusch et al., 2021).

While in the two examples above, learners or teachers are responsible for making decisions, a third option is to leave the decision about adapting the learning environment to the learning environment itself. The idea of this approach of adaptivity in learning contexts is to provide learners with the exact learning experience and support that learners need in a particular situation to successfully achieve intended learning goals (Plass and Pawar, 2020).

By adapting learning environments and the therein contained support structures to the learners' needs, personalized learning becomes possible (Bernacki et al., 2021). Reviews show that personalized learning in adaptive learning environments can have a positive impact on student learning (see Aleven et al., 2016; Bernacki et al., 2021; Ninaus and Nebel, 2021). However, more specific questions, such as which aspects of learning environments and according to which variables should be adapted to in order to foster learning remains largely unresolved.

Regarding adjustments of learning environments, macro-level and micro-level adaptivity can be distinguished (Plass and Pawar, 2020). On the one hand, macro-level adaptivity refers to adjustments regarding general categories of the wider learning context like the provision of suggestions for suitable learning material or courses based on the aggregation of events in learning environments (Sevarac et al., 2012; Mah and Ifenthaler, 2018). On the other hand, micro-level adaptivity focuses on currently processed learning tasks and thus on adapting the learning environment to the learner's needs just-in-time (Plass and Pawar, 2020). If we consider the question of how micro-adaptivity can be established in learning environments, feedback approaches (Hattie and Timperley, 2007) and scaffolding approaches (Belland et al., 2017) stand out.

Especially for complex learning tasks, providing feedback on process or self-regulation level is necessary to master the necessary steps for solving a problem or to effectively monitor task performance (Wisniewski et al., 2020). Adaptive feedback might be especially promising on process or self-regulation level to develop an understanding of the current state of knowledge and identify the differences to an optimal state of knowledge. Further, adaptive feedback can feed back flawed task processing just in time (Narciss et al., 2014; Bimba et al., 2017). While some of this ideas have been tested in the context of intelligent tutoring systems, which are based on logfiles and closed-end questions (Graesser et al., 2018), AI-based methods can also provide a merit when complex tasks require students to write open text answers. AI-based approaches like Natural Language Processing (Manning and Schütze, 2005) can automatically analyze written text and allow for adaptively activating different feedback elements or different solutions based on the students' answers (Zhu et al., 2017, 2020; Sailer et al., 2022).

Besides adaptive feedback, different forms of adaptive scaffolding are promising in the context of AI. The basic idea of scaffolding is to support learners in their problem solving, thus promoting their acquisition of knowledge and skills (Belland et al., 2017). As the need for support can vary between and within learners during task processing, the idea of adaptive scaffolding is to provide students with the support they need in specific situations at a specific time (Radkowitsch et al., 2021). Cognitive, meta-cognitive (see Belland et al., 2017), socio-cognitive (see Radkowitsch et al., 2020), and affective-motivational scaffolds (see Schrader and Bastiaens, 2012) can profit from the use of AI as they can be precisely faded in or out depending on learners needs. However, also other types of adaptive scaffolds that address the complexity of the learning environment or the salience of particular aspects of a learning environment or a learning task might profit from the use of AI. This form of indirect support can be referred to as representational scaffolding (Fischer et al., 2022). Representational scaffolds can be used to systematically vary the complexity of the learning environment and the salience of its aspects relevant to learning (Stadler et al., 2019b; Chernikova et al., 2020) in order to enable learners to solve problems according to their respective levels of knowledge and skills (e.g., Stadler et al., 2019a).

Closing the loop

As highlighted above, AI-supported learning systems rely on decisions made in several steps along the proposed loop (see **Figure 1**). In a nutshell, user data is recorded, from which relevant data can be pre-selected using theoretical (human decision) as well as data-driven (machine) selection processes. In a next step, relevant patterns in data are detected by specifically selected ML algorithms. Based on successful pattern detection, suggestions regarding adaptations of the learning environment to the learners needs are provided to teachers or learners or decisions about adaptations are directly executed by the system. Finally, the result of this personalization affects the users' learning process, which will be reflected in the data. This new user data can be used to refine the overall process, for instance, by identifying patterns that indicate potential improvements of the user and their learning process, which in turn will affect personalization procedures.

The proposed closed-loop highlights the complexity of AI-supported learning systems. Some of the manifold decisions described in the different steps can be automated using digital technologies and AI. There is also evidence that users prefer judgments from algorithms instead of judgments of people, despite blindness to the algorithm's process (Logg et al., 2019). However, in many respects, human decisions are essential in the process (see Baker, 2016; Ritter et al., 2016; Holstein et al., 2017, 2020) and require expertise and perspectives

from various disciplines (e.g., Sailer et al., 2022). In this perspective article, we want to emphasize the crucial role of human decisions in the design and implementation process of AI in education. Accordingly, we suggest striving for hybrid solutions by balancing the process of human- and AI-driven decisions and mutual monitoring of these decisions, which is in line with current discussions and frameworks on AI use in education (see Holstein et al., 2020; Molenaar, 2021) and beyond such as medicine (e.g., for detecting tumors, Topol, 2019) and autonomous driving (Awad et al., 2018; Ning et al., 2022) where AI is already more established. In these latter domains, AI technology is still mainly used to support or assist humans but has not replaced them. In fact, intricate moral decisions (Awad et al., 2018) and discussions revolving around bias, transparency, privacy, and accuracy are at the center of AI applications in these domains (Topol, 2019), which will also increasingly accompany the use and implementation of AI in education (for a detailed discussion see Akgun and Greenhow, 2021). Furthermore, as learning is a highly complex process, we would argue that in education, we still have a very long way to go to utilize AI in a balanced way, and – similar to medicine and autonomous driving – hybrid solutions will be dominant. The boundaries between AI and human decision-making, however, will definitely fluctuate (see Molenaar, 2021).

In the context of education, we believe that AI will change or shape the responsibilities and tasks of the different stakeholders involved in the educational process (see Molenaar, 2021 for a more detailed description of the teachers' role in hybrid human-AI systems), which might differ across learning domains, contexts, situations. Accordingly, we want to emphasize the critical role of human decisions in high stake situations. Let us think back, for instance, at the example in the beginning using the AI-supported LMS that drew the wrong conclusions and thus provided you with an incorrect adaptation. Let us add to this a situation with more serious consequences: It has been argued that AI-supported systems might be useful for grading (e.g., Rus et al., 2013; Timms, 2016; Chen et al., 2020), selection of promising candidates for a job (e.g., Black and van Esch, 2020), or even for healthcare decisions (e.g., Pakdemirli, 2019). In fact, AI-supported systems can be a massive support for all those circumstances, but we need to be aware that those systems are not 100% accurate but can commit errors.

We can contextualize these decisions or erroneous conclusions, for instance, within statistician hypothesis testing and differ between type I (e.g., the system classifies a pupil to be not ready for higher secondary education when they actually are) and type II errors (e.g., the system predicts someone to pass the class when indeed the person will fail). Type I or type II errors can have very different consequences, and accordingly, one has to decide on a case-by-case basis how much decision-making power is given to an AI. In most cases, a hybrid decision-making process will probably be most correct and fair. In particular, AI in education might be used to support decision making, i.e., basing the decision process on insights or even recommendations provided by the AI and your own experience, impressions, and conclusions (Ritter et al., 2016; Holstein et al., 2020). While neither the AI nor the humans involved will always make correct decisions, the decision-making process can be improved by taking both sides into account. For instance, when an AI comes to the same conclusions as a teacher, correct conclusions are more likely. In contrast, disagreements between AI and the teacher might shed led on potential erroneous conclusions that otherwise would have remained hidden. We hope that by showing the steps of an AI-supported system, we demonstrated that humans can have a crucial role at many stages in this process and that we can use AI to support our capacities.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

Both authors have an equal contribution during the process of conceptualizing and writing this perspective article. Both authors approved the submitted version.

Acknowledgments

The authors acknowledge the financial support by the University of Graz.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akgun, S., and Greenhow, C. (2021). Artificial intelligence in education: addressing ethical challenges in K-12 settings. *AI Ethics* [Epub ahead of print]. doi: 10.1007/s43681-021-00096-7

Aleven, V., McLaughlin, E. A., Glenn, R. A., and Koedinger, K. R. (2016). "Instruction based on adaptive learning technologies," in *Handbook of research on learning and instruction*, eds R. E. Mayer and P. A. Alexander (Abingdon: Routledge), 522–560. doi: 10.4324/9781315736419.ch24

Appel, T., Gerjets, P., Hoffman, S., Moeller, K., Ninaus, M., Scharinger, C., et al. (2021). Cross-task and cross-participant classification of cognitive load in an emergency simulation game. *IEEE Trans. Affective Comput*. doi: 10.1109/TAFFC.2021.3098237 [Epub ahead of print].

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature* 563, 59–64. doi: 10.1038/s41586-018-0637-6

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., et al. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *Int. J. Educ. Technol. High Educ.* 17:13. doi: 10.1186/s41239-020-00187-1

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *Int. J. Artif. Intell. Educ.* 26, 600–614. doi: 10.1007/s40593-016-0105-0

Belland, B. R., Walker, A. E., Kim, N. J., and Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. *Rev. Educ. Res.* 87, 309–344. doi: 10.3102/0034654316670999

Bernacki, M. L., Greene, M. J., and Lobczowski, N. G. (2021). A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? *Educ. Psychol. Rev*. 33, 1675–1715. doi: 10.1007/s10648-021-09615-8

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., and Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: a review. *Adapt. Behav.* 25, 217–234. doi: 10.1177/1059712317727590

Black, J. S., and van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Bus. Horizons* 63, 215–226. doi: 10.1016/j.bushor.2019.12.001

Brandl, L., Richters, C., Radkowitsch, A., Fischer, M. R., Schmidmaier, R., Fischer, F., et al. (2021). Simulation-based learning of complex skills: predicting performance with theoretically derived process features. *Psychol. Test Assess. Model.* 63, 542–560.

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access* 8, 75264–75278. doi: 10.1109/ACCESS.2020.2988510

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., Fischer, F., et al. (2020). Facilitating diagnostic competences in higher education—a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Clow, D. (2012). "The learning analytics cycle: closing the loop effectively," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, New York, NY.

Di Mitri, D., Schneider, J., Specht, M., and Drachsler, H. (2018). From signals to knowledge: a conceptual model for multimodal learning analytics. *J. Comput. Assist. Learn.* 34, 338–349. doi: 10.1111/jcal.12288

Drachsler, H., and Greller, W. (2016). "Privacy and Analytics – it's a DELICATE Issue A Checklist for Trusted Learning Analytics," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, New York, NY.

Fischer, F., Bauer, E., Seidel, T., Schmidmaier, R., Radkowitsch, A., Neuhaus, B., et al. (2022). Representational scaffolding in digital simulations – learning professional practices in higher education. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/bf92d

Gal, R., May, A. M., van Overmeeren, E. J., Simons, M., and Monninkhof, E. M. (2018). The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. *Sports Med. Open* 4:42. doi: 10.1186/s40798-018-0157-9

Graesser, A. C., Hu, X., and Sottilare, R. (2018). Intelligent tutoring systems. *Int. Handb. Learn. Sci.* 246–255. doi: 10.4324/9781315617572

Hakimi, L., Eynon, R., and Murphy, V. A. (2021). The ethics of using digital trace data in education: a thematic review of the research landscape. *Rev. Educ. Res.* 91, 671–717. doi: 10.3102/00346543211020116

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., et al. (2021). Machine learning for the educational sciences. *Rev. Educ.* 9:e3310. doi: 10.1002/rev3.3310

Holstein, K., Aleven, V., and Rummel, N. (2020). "A Conceptual Framework for Human–AI Hybrid Adaptivity in Education," in *Artificial Intelligence in Education Lecture Notes in Computer Science*, eds I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán (Cham: Springer International Publishing), 240–254. doi: 10.1007/978-3-030-52237-7_20

Holstein, K., McLaren, B. M., and Aleven, V. (2017). "Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, New York, NY.

Huijsmans, M. D. E., Kleemans, T., van der Ven, S. H. G., and Kroesbergen, E. H. (2020). The relevance of subtyping children with mathematical learning disabilities. *Res. Dev. Disabil.* 104:103704. doi: 10.1016/j.ridd.2020.103704

Kober, S. E., Ninaus, M., Friedrich, E. V. C., and Scherer, R. (2018). "BCI and Games: Playful, Experience-Oriented Learning by Vivid Feedback?," in *Brain–Computer Interfaces Handbook: Technological and Theoretical Advances*, eds C. S. Nam, A. Nijholt, and F. Lotte (Boca Raton, FL: CRC Press - Taylor & Francis Group).

Li, Q., Baker, R., and Warschauer, M. (2020). Using clickstream data to measure, understand, and support self-regulated learning in online courses. *Internet High. Educ.* 45:100727. doi: 10.1016/j.iheduc.2020.100727

Liarokapis, F., Debattista, K., Vourvopoulos, A., Petridis, P., and Ene, A. (2014). Comparing interaction techniques for serious games through brain-computer interfaces: a user perception evaluation study. *Entertain. Comput.* 5, 391–399. doi: 10.1016/j.entcom.2014.10.004

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005

Mah, D.-K., and Ifenthaler, D. (2018). Students' perceptions toward academic competencies: the case of German first-year students. *Issues Educ. Res.* 28, 120–137. doi: 10.3316/informit.437867582603162

Mangaroska, K., and Giannakos, M. (2019). Learning analytics for learning design: a systematic literature review of analytics-driven design to enhance learning. *IEEE Trans. Learn. Technol.* 12, 516–534. doi: 10.1109/TLT.2018.2868673

Manning, C., and Schütze, H. (2005). *Foundations of Statistical Natural Language Processing*, 8th Edn. Cambridge, MA: MIT Press.

Molenaar, I. (2021). *Personalisation of learning: Towards hybrid human-AI learning technologies*. Paris: OECD Publishing.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., et al. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Comput. Educ.* 71, 56–76. doi: 10.1016/j.compedu.2013.09.011

Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., et al. (2019). Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data. *Comput. Educ.* 142:103641. doi: 10.1016/j.compedu.2019.103641

Ninaus, M., and Nebel, S. (2021). A systematic literature review of analytics for adaptivity within educational video games. *Front. Educ.* 5:611072. doi: 10.3389/feduc.2020.611072

Ninaus, M., Witte, M., Kober, S. E., Friedrich, E. V. C., Kurzmann, J., Hartsuiker, E., et al. (2013). "Neurofeedback and Serious Games," in *Psychology, Pedagogy, and Assessment in Serious Games*, eds T. M. Connolly, E. Boyle, T. Hainey, G. Baxter, and P. Moreno-ger (Hershey, PA: IGI Global), 82–110. doi: 10.4018/978-1-4666-4773-2.ch005

Ning, H., Yin, R., Ullah, A., and Shi, F. (2022). A survey on hybrid human-artificial intelligence for autonomous driving. *IEEE Trans. Intell. Transport. Syst.* 23, 6011–6026. doi: 10.1109/TITS.2021.3074695

Pakdemirli, E. (2019). Artificial intelligence in radiology: friend or foe? Where are we now and where are we heading? *Acta Radiol. Open* 8:205846011983022. doi: 10.1177/2058460119830222

Plass, J. L., and Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *J. Res. Technol. Educ.* 52, 275–300. doi: 10.1080/15391523.2020.1719943

Radkowitsch, A., Sailer, M., Schmidmaier, R., Fischer, M. R., and Fischer, F. (2021). Learning to diagnose collaboratively – Effects of adaptive collaboration

scripts in agent-based medical simulations. *Learn. Instr.* 75:101487. doi: 10.1016/j.learninstruc.2021.101487

Radkowitsch, A., Vogel, F., and Fischer, F. (2020). Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts. *Intern. J. Comput. Support. Collab. Learn.* 15, 5–47. doi: 10.1007/s11412-020-09316-4

Ritter, S., Yudelson, M., Fancsali, S. E., and Berman, S. R. (2016). "How Mastery Learning Works at Scale," in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, Edinburgh.

Rus, V., D'Mello, S., Hu, X., and Graesser, A. (2013). Recent advances in conversational intelligent tutoring systems. *AIMag* 34, 42–54. doi: 10.1609/aimag.v34i3.2485

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., et al. (2022). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learn. Instr.* [Preprint]. doi: 10.1016/j.learninstruc.2022.101620

Schrader, C., and Bastiaens, T. (2012). Learning in educational computer games for novices: the impact of support provision types on virtual presence, cognitive load, and learning outcomes. *IRRODL* 13:206. doi: 10.19173/irrodl.v13i3.1166

Sevarac, Z., Devedzic, V., and Jovanovic, J. (2012). Adaptive neuro-fuzzy pedagogical recommender. *Expert Syst. Applic.* 39, 9797–9806. doi: 10.1016/j.eswa.2012.02.174

Stadler, M., Niepel, C., and Greiff, S. (2019b). Differentiating between static and complex problems: a theoretical framework and its empirical validation. *Intelligence* 72, 1–12. doi: 10.1016/j.intell.2018.11.003

Stadler, M., Fischer, F., and Greiff, S. (2019a). Taking a closer look: an exploratory analysis of successful and unsuccessful strategy use in complex problems. *Front. Psychol.* 10:777. doi: 10.3389/fpsyg.2019.00777

Timms, M. J. (2016). Letting artificial intelligence in education out of the box: educational cobots and smart classrooms. *Int. J. Artif. Intell. Educ.* 26, 701–712. doi: 10.1007/s40593-016-0095-y

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Wiedbusch, M. D., Kite, V., Yang, X., Park, S., Chi, M., Taub, M., et al. (2021). A theoretical and evidence-based conceptual design of metadash: an intelligent teacher dashboard to support teachers' decision making and students' self-regulated learning. *Front. Educ.* 6:570229. doi: 10.3389/feduc.2021.570229

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: a meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3389/fpsyg.2019.03087

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., and Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *Int. J. Sci. Educ.* 39, 1648–1668. doi: 10.1080/09500693.2017.1347303

Zhu, M., Liu, O. L., and Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Comput. Educ.* 143:103668. doi: 10.1016/j.compedu.2019.103668

frontiers | Frontiers in Artificial Intelligence

TYPE Mini Review
PUBLISHED 13 September 2022
DOI 10.3389/frai.2022.903051

OPEN ACCESS

EDITED BY
Mario Allegra,
National Research Council (CNR), Italy

REVIEWED BY
Malinka Ivanova,
Technical University of Sofia, Bulgaria
Vincenza Benigno,
National Research Council (CNR), Italy

*CORRESPONDENCE
Tanya Chichekian
tanya.chichekian@usherbrooke.ca

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 23 March 2022
ACCEPTED 22 August 2022
PUBLISHED 13 September 2022

CITATION
Chichekian T and Benteux B (2022) The potential of learning with (and not from) artificial intelligence in education. *Front. Artif. Intell.* 5:903051. doi: 10.3389/frai.2022.903051

The potential of learning with (and not from) artificial intelligence in education

Tanya Chichekian* and Bérenger Benteux

Faculty of Education, Université de Sherbrooke, Longueuil, QC, Canada

AI-powered technologies are increasingly being developed for educational purposes to contribute to students' academic performance and overall better learning outcomes. This exploratory review uses the PRISMA approach to describe how the effectiveness of AI-driven technologies is being measured, as well as the roles attributed to teachers, and the theoretical and practical contributions derived from the interventions. Findings from 48 articles highlighted that learning outcomes were more aligned with the optimization of AI systems, mostly nested in a computer science perspective, and did not consider teachers in an active role in the research. Most studies proved to be atheoretical and practical contributions were limited to enhancing the design of the AI system. We discuss the importance of developing complementary research designs for AI-powered tools to be integrated optimally into education.

KEYWORDS

artificial intelligence, intelligent tutoring system, learning, performance, education

Introduction

In the last decade, there has been a surge of educational research about how to effectively integrate technology in classrooms, with a focus on providing digital experiences that improve students' academic performance. With recent movements regarding the use of artificial intelligence (AI) as the leading medium by which we engage students in scholarly tasks (Roll and Wylie, 2016), rethinking how to design such technology is imperative if the intent is to facilitate the learning processes that lead to the achievement of learning objectives and, ultimately, to optimal functioning in education.

AI is defined by Popenici and Kerr (2017) as "computing systems that can engage in human-like processes such as learning, adapting, synthesizing, self-correction and use of data for complex processing tasks" (p. 2). These systems, which are displayed in various forms ranging from Internet search engines to smartphone applications, are shaping new teaching and learning educational contexts (Pedró et al., 2019). Generally, educational technologies driven by AI-powered algorithms are referred to as Intelligent Tutoring System (ITS) and try to replicate human tutor interactions (VanLehn, 2011) through a pedagogical agent by providing timely feedback and guidance to students (Kay, 2012).

However, while ITSs have made advancements in the field of education, specifically in online environments or in computer labs, it remains unclear as to how their effectiveness can be interpreted or translated regarding the quality of students' learning outcomes (Pedró et al., 2019). This is partly due to the minimal evidence and support for wider adoption of the term "learning" on the part of the AIED community, and even less attention attributed to developing a well-defined role for teachers implementing these technologies. The latter is reflected in AIED research being published mostly in specialized journals and conference proceedings, which rarely become visible to educational researchers and only include limited educational perspectives in line with these technological developments (Zawacki-Richter et al., 2019). Although there is some strength with AIED and ITS conferences providing opportunities for the cross-fertilization of approaches, techniques, and ideas stemming from multidisciplinary research fields, it also creates a massive challenge for the AIED community in terms of communicating successfully both within the field and beyond, particularly with key actors in the wider education community. Tuomi (2018) also stated that there is a high chance that the way current AIED systems are being designed and developed is far from the learning outcomes learning scientists and teachers are expecting from these tools, especially if most AIED research has a weak connection to theoretical and pedagogical perspectives and is more aligned as a system of inputs and outputs.

The current exploratory review's purpose was to interpret findings from AIED research using a pedagogical perspective. Such an investigation is important for the following reasons. First, practical implications need to be considered in the field of education if certain conditions are to be fulfilled and resources reinvested to facilitate pedagogical activities. This is essential when examining whether the AIED field has the potential to be impactful in authentic situations. Such practical implications are also helpful for decision-makers in determining adequate funding policies for AIED research and projects. Second, given the growing trend in the number of publications in AIED (Chen et al., 2020a,b), interest in the field is in expansion. As such, the publication sources and conference venues play a major role in helping the educational community identify relevant information and findings that can be reflected in the progress and advancement of this field in educational settings. Third, it guides individuals from different disciplines to be exposed, to understand, and analyze the use of AI-driven technologies from multiple perspectives and thus visualize innovative ways of adapting them for educational purposes.

Accordingly, this review aims to answer the following research questions:

- RQ1: How is the effectiveness of an ITS measured in AIED research?
- RQ2: To what extent does AIED research contribute to the field of education?

This review contributes to the research field by enabling the educational community to understand the relationship between students' learning gains and the role of the ITS. Furthermore, it provides a knowledge base from which educational and computer science researchers can extrapolate to build, design, and collaborate on projects that are suitable for scaling up.

State-of-the-art in AIED research

According to a meta-analysis conducted by Ma et al. (2014), an ITS is composed of four elements: (1) an interface that communicates with the learner by presenting information, asking questions, assigning learning tasks, providing feedback, and answering questions posed by students, (2) a domain model that represents the knowledge intended for the student to learn, (3) a student model that represents relevant aspects of the student's knowledge or psychological state determined by the student's responses to questions or other interactions with the interface, and (4) a tutoring model that adapts instructional strategies based on the needs of the learners. On a cognitive level, ITSs have facilitated students' learning processes during homework and practice exercises in the absence of a teacher or a tutor (VanLehn, 2011). Given that their use has often led to significantly higher achievement outcomes compared to other modes of instruction (Ma et al., 2014), they are often considered one of the resources in educators' toolboxes (Steenbergen-Hu and Cooper, 2013). In terms of supporting students' learning processes, ITSs seem to be most impactful on metacognitive strategies by prompting students to apply self-regulation skills and monitor their progress when learning (Bouchet et al., 2016). For example, Verginis et al. (2011) showed that the use of an open-learner model guided previously disengaged online students toward re-engagement and, ultimately, to improved post-test performance. Similarly, Arroyo et al. (2014) provided evidence of the positive impact that learning companions had on the improvement of low-achieving students' affective states and their motivation. It seems ITSs occupy an important and complementary place in learning and as a supplement to teachers' instruction.

Impact on learning

In the last couple of years, numerous studies have started demonstrating how the use and impact of digital technologies and ITSs are directly related to the extent to which the technology itself is responsible for the observed increases in students' academic performance. Results from a meta-analysis (Schroeder et al., 2013) showed how pedagogical agents (PAs) had a small but significant positive effect on learning ($g = 0.19$, $p < 0.05$) among K-12 students compared to those who did not interact with PAs. Their learning gains were proportionally

higher compared to collaborative interactions with other types of traditional, closed-ended, and teacher-led interactions or with non-ITS computer-based instruction (Harley et al., 2018). Arroyo et al. (2014) also demonstrated how students in these types of collaborative activities not only displayed an increase in learning gains but also passed standardized tests more frequently (92%) compared to a control group (79%) or to students who did not interact with any tutor (76%).

According to Steenbergen-Hu and Cooper (2013, 2014), certain variables have also been tested for moderating the significance of the effectiveness of an ITS. These include comparison conditions (e.g., traditional classroom instruction), type of ITS, subject matter, year of study, teacher involvement, assessment type, schooling level, length of the intervention, degree of implementation, students' prior knowledge, sample size, research design, self-regulation skills, and academic motivation. Specifically, ITSs produced the most significant impact depending on: (1) the year of study, (2) teacher involvement, and (3) the use of embedded assessments. It was rarely reported how process variables might help to explain observed effects or a lack thereof (Winne and Baker, 2013). Ma et al. (2014) indicated that whenever a process variable was reported in a study, it was often only meaningful in the context of the learning task. Therefore, when referring to the outperformance of an ITS compared to other methods of computer-based instruction, the effect at the level of computer-student interaction was rarely considered. Nevertheless, the use of ITSs to increase academic achievement was significant regardless of the context in which it was used. However, despite its effectiveness as a learning tool, the emergence and rapid growth of technology in education have resulted in a rushed deployment with not enough time to analyze how learning should be measured with the assistance of AI nor the extent to which teachers should implement these AI-driven learning experiences in the curriculum (Pedró et al., 2019).

AI-driven learning experiences

Research and development on AIEd is still a young field in which the advancement of knowledge has the potential to make significant contributions to the learning sciences. Steenbergen-Hu and Cooper (2014) suggested various pedagogical hypotheses to move in such a direction such as experimentally adjusting the type of instruction and the frequency of feedback to optimize instruction and ITS equitably and meet the needs of different learners. Examples of such design strategies have thus far resided in the Computer-Human Interaction field. For example, Positive Technologies (Riva et al., 2012) applied templates from positive psychology to improve the technology's affective quality as well as promote students' engagement and connectedness with the content. Hassenzahl (2010) proposed an experiential approach for design to explore what matters to humans, what is needed to make technology more meaningful, and how to uncover "experience patterns" in human activities. Similarly, Positive Design (Desmet and Pohlmeyer, 2013), a framework for wellbeing, focused on how the design of any artifact or product might foster flourishing. Finally, as part of Positive Computing, Calvo and Peters (2014) provided leverage for a design supportive of wellbeing as well as its determinants. Although each of these frameworks provides some version of the core elements that are foundational to the learning sciences, these models remain at a distance from the field of education. Many articles about educational technology remain atheoretical (Hew et al., 2019) and lack focus on pedagogical perspectives (Bartolomé et al., 2018). To better understand, empirically evaluate, and design learning experiences about the impact of AI-driven technologies on students' academic success, as well as on certain psychological aspects that play a role in the learning process such as their motivation, we need to anchor them in conceptual or theoretical frameworks that take origin at the intersections of education, psychology, and computer science. One example is the Motivation, Engagement, and Thriving in User Experience model that was based on the self-determination theory (SDT, Deci and Ryans, 2002) to assess psychological needs in five different but interdependent contexts: at the point of technology adoption, during interaction with the interface, as a result of engagement with technology-specific tasks, as part of the technology-supported behavior, and as part of an individual's life overall. In addition to predicting the impact on motivation and sustained engagement with technology, the SDT can also serve as a basis to measure educational or other domain-specific outcomes, thus making it an ideal framework on which to build an understanding of common goals within technology projects.

Research method

We searched the literature in the ERIC, PsycINFO, and Education Source databases as they contained the most publications regarding educational research. We used a combination of terms from the AIED and education fields such as "artificial intelligence," "intelligent tutoring systems," "natural language processing," "student*," and "learn*". Additionally, we included synonyms found in the search databases' thesaurus that related to the term impact such as "impact*," "effect*," "outcome*," "consequence*," and "eval*". More specifically, we searched the mentioned databases with the keywords "artificial intelligence" or "intelligent tutor* systems" in the topic, then we used the connector "AND" to combine these results with the keywords "student*" or "learn*" and in a third step we combined these results, using the connector "AND", with our keywords about impacts, namely "impact*" or "effect*" or "outcome*" or "consequence*" or "eval*". This search resulted in a total of 479 articles (386 articles in ERIC, 79 in PsycINFO,

and 32 in Education Source) which we imported to Zotero, a reference management system.

To begin the screening process, we first checked for duplicates which resulted in the deletion of 41 articles. We then scanned the remaining articles to decide if they met the following inclusion criteria:

- Evaluated the effects of AI on learning;
- Published in peer-reviewed journals;
- Took place between 2009 and 2019;
- Published in English.

A study needed to meet all our criteria to be included in the review. After deleting duplicates and applying these selection criteria to the remaining 438 publications, we narrowed down the review to 48 articles (see Figure 1). The most common reasons for which studies did not qualify for inclusion were that they focused primarily on the description of the design or development of a system, they addressed non-AI-powered educational technologies, and they evaluated systems' (not learner') outcomes. We coded and analyzed studies based on the following elements: the workplace and departments where the authors of the selected articles worked, theoretical framework, teachers' attributed role in the study, learning outcomes, as well as theoretical and practical contributions. In the following section, we present the current achievements in AIED followed by a summary of the findings from this literature search.

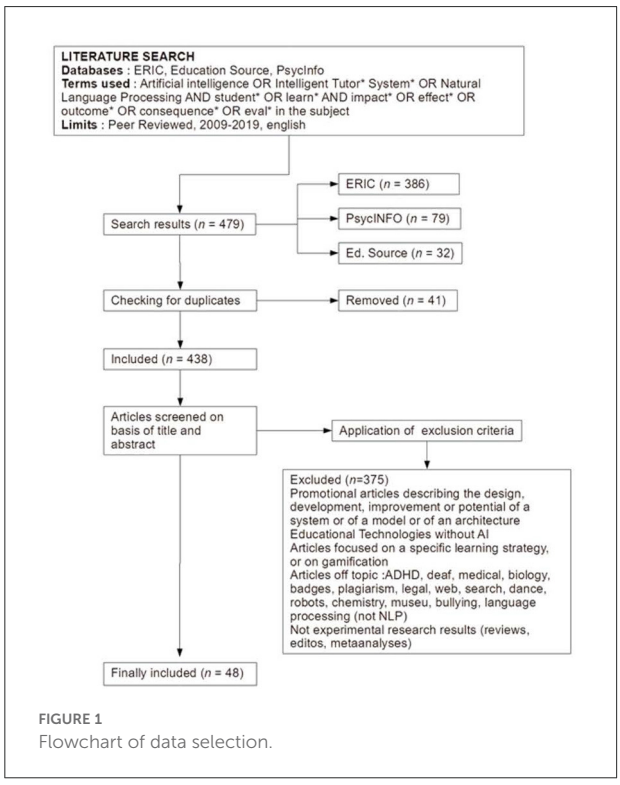


FIGURE 1
Flowchart of data selection.

Findings, analysis, and discussion

The effectiveness of an ITS as measured in AIED research

Findings (see Supplementary Table 1) indicated that the effectiveness of an ITS was assessed by measuring students' learning gains, either as a difference between a pre and post-test [n =15 (31%)], as a perception of student learning [n = 19 (40%)], as a level of interaction with a learner during an activity, n = 6 (13%), or through standardized measurements such as national tests [n = 5 (10%)] or academic performance [n = 8 (17%)]. These results are in line with other meta-analytic reviews such as those from Kulik and Fletcher (2016) demonstrating the effectiveness of ITSs as instructional tools. Compared to students in conventional classes, those who received intelligent tutoring outperformed the others 92% of the time and this improved performance was significant enough 78% of the time. Moreover, the effectiveness of ITSs at times even surpassed other forms of computer or human tutoring.

These results are not surprising given that three meta-analytic reviews (Ma et al., 2014; Steenbergen-Hu and Cooper, 2014; Kulik, 2015) had also revealed the effectiveness of the ITS-related learning to be reflective of improved test score measurements. Our findings also concur with that of Guo et al. (2021) who advocated that despite high-level research in AIED, many were repetitive with few innovative breakthroughs in recent years. From an educational stance, this implies that the effects and functions that ITSs are seeking to achieve remain limited and, consequently, void of guidelines emanating from more robust theoretical frameworks nested in the learning sciences.

AIED's contribution to the field of education

To examine the theoretical contributions of the selected studies to the AIED research community, we first noted the industry in which the authors worked. Of the 169 authors of the 48 articles, 78% (n = 132) were professors in a university department: 15% (n = 26) from education, 4% (n = 7) from educational psychology, 32% (n = 54) computer science, 12% (n = 20) from engineering, 11% (n = 18) from psychology, and 4% (n = 7) from other departments. In addition, 12% (n = 20) of the authors worked in a university, but not as a professor, and 10% (n = 17) were not from academia. Next, our review revealed that only n = 10 (6%) studies referred to a theoretical framework that supported their research study of which n = 6 originated from the field of educational psychology and n = 4 from pedagogy. Finally, of the 10 articles referring to a theoretical framework, n = 6 (4%) also mentioned a theoretical contribution of their study findings to the field (n = 6

in educational psychology, $n = 4$ in pedagogy, and $n = 1$ in cognitive psychology), either as a replication of previous research ($n = 4$) or as an extension to current knowledge ($n = 2$). In terms of the practical contributions associated with each study, $n = 39$ (81%) studies targeted the optimization of the ITS's performance and $n = 8$ contributed to improving the design of an ITS. Only one article had no practical contribution as its goal was to make a theoretical contribution and confirm previous research in the field. The significant gap between the theoretical and practical contributions is aligned with the focus on online learning, especially during the pandemic with an exponential increase in the utilization of AI-powered educational technology tools (Chaudhry and Kazim, 2022). A lot more work needs to be done on outlining the theoretical contributions of AIED as we move forward with a growing number of AI-powered educational technology that has the potential of producing a long-lasting educational and psychological impact on learners and teachers equally.

Overall, these findings demonstrate that the field of AIED seems to be targeting outcomes related more to the optimization of AI systems compared to the quality of learning itself. Moreover, most of the studies we reviewed only evaluated the impacts of these AI-powered technologies from a computer science perspective. Rarely were the studies framed and conceived as research contributing to a theoretical question about the relationship between ITS and learning outcomes. This is in line with past findings revealing very little evidence for the advancement of a pedagogical perspective and learning theories in AIED research (Bartolomé et al., 2018). To develop a complementary research design embedded within an educational framework (Pedro, 2019), integrating interdisciplinary perspectives about how to use AI for learning in educational settings is a future avenue worth exploring. It seems there is still substantial room to adopt a more participatory approach with educators if the field of AIED is to produce a critical reflection regarding the pedagogical and ethical implications of implementing AI applications in classrooms and, more importantly, to contribute to the advancement of learning theories with an appropriate and aligned conceptual or theoretical framework.

Conclusion

This exploratory review highlighted that the purpose of most educational research with AI-driven technologies was to demonstrate the effectiveness of an ITS by measuring students' academic performance. Although recent studies have shown how these technologies also contribute to overall better learning outcomes among students (Laanpere et al., 2014; Luckin et al., 2016), very few have been implemented as applications in classrooms. To capitalize on students' optimal learning (Ryan and Deci, 2017), in addition to academic performance, positive learning experiences need to be designed that consider students' interactions with AI, including the maintenance of a certain level of motivation and engagement (Niemic and Ryan, 2009; Peters et al., 2018), as well as a well-defined role for the classroom teacher. With a more proactive role assigned to classroom teachers involved in collaborative research, the need to integrate their perspectives in AI-driven educational technological developments would be integral in understanding student learning in a sociotechnical approach. Combining a technical system and a classroom culture requires different levels of adaptability given that an ITS needs to be designed and integrated into both the students' learning and the teacher's instruction. Perhaps the next challenge for the AIED community is to determine a more equitable division of labor between the roles of the teacher and of the intelligent tutoring system, both of which support students with instructions, tasks, and feedback. Self-adaptive systems could enable a transformation in educational practice (Tuomi, 2018), but the challenge remains in deciding whether the intelligent tutoring system or the teaching activities should be re-designed and re-aligned with the other.

Author contributions

TC contributed to conception and design of the study and the other sections of the manuscript. BB organized the database, performed the search, selection of studies for the review, and wrote the results. All authors contributed to manuscript revision.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.903051/full#supplementary-material

References

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., and Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *Int. J. Artif. Intellig. Educ.* 24, 387–426. doi: 10.1007/s40593-014-0023-y

Bartolomé, A., Castañeda, L., and Adell, J. (2018). Personalisation in educational technology: the absence of underlying pedagogies. *Int. J. Educ. Technol. Higher Educ.* 15, 14. doi: 10.1186/s41239-018-0095-0

Bouchet, F., Harley, J. M., and Azevedo, R. (2016). "Can adaptive pedagogical agents' prompting strategies improve students' learning and self-regulation?", in *Intelligent Tutoring Systems*, eds. A. Micarelli, J. Stamper, and K. Panourgia (Cham: Springer International Publishing), 368–374.

Calvo, R. A., and Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. Ambrisge, MA: MIT Press.

Chaudhry, M.A., and Kazim, E. (2022). Artificial intelligence in education (AIEd): a high-level academic and industry note 2021. *AI Ethics* 2, 157–165. doi: 10.1007/s43681-021-00074-z

Chen, X., Xie, H., and Hwang, G.-J. (2020a). A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers. *Comput. Educ. Artif. Intellig.* 1, 100005. doi: 10.1016/j.caeai.2020.100005

Chen, X., Xie, H., Zou, D., and Hwang, G.-J. (2020b). Application and theory gaps during the rise of Artificial Intelligence in Education. *Comput. Educ. Artif. Intellig.* 1, 100002. doi: 10.1016/j.caeai.2020.100002

Desmet, P. M. A., and Pohlmeyer, A. E. (2013). Positive design: an introduction to design for subjective well-being. *Int. J. Design* 7, 5–19.

Guo, L., Wang, D., Gu, F., Li, Y., Wang, Y., and Zhou, R. (2021). Evolution and trends in intelligent tutoring systems research: a multidisciplinary and scientometric view. *Asia Pacific Educ. Rev.* 22, 441–461. doi: 10.1007/s12564-021-09697-7

Harley, J. M., Taub, M., Azevedo, R., and Bouchet, F. (2018). "Let's set up some subgoals": understanding human-pedagogical agent collaborations and their implications for learning and prompt and feedback compliance. *IEEE Trans. Learn. Technol.* 11, 629. doi: 10.1109/TLT.2017.2756629

Hassenzahl, M. (2010). Experience design: technology for all the right reasons. *Synth. Lect. Hum. Cent. Inform.* 3, 1–95. doi: 10.1007/978-3-031-02191-6

Hew, K. F., Lan, M., Tang, Y., Jia, C., and Lo, C. K. (2019). Where is the "theory" within the field of educational technology research? *Br. J. Educ. Technol.* 50, 956–971. doi: 10.1111/bjet.12770

Kay, J. (2012). AI and education: grand challenges. *IEEE Intellig. Syst.* 27, 66–69. doi: 10.1109/MIS.2012.92

Kulik, J. A., and Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* 86, 42–78. doi: 10.3102/0034654315581420

Laanpere, M., Pata, K., Normak, P., and Põldoja, H. (2014). Pedagogy-driven design of digital learning ecosystems. *Comput. Sci. Inform. Syst.* 11, 419–442. doi: 10.2298/CSIS121204015L

Luckin, R., Holmes, W., Griffiths, M., and Forcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in Education*. London: UCL Knowledge Lab. Available online at: https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/innovation/Intelligence-Unleashed-Publication.pdf (accessed April, 2022).

Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* 106, 901–918. doi: 10.1037/a0037123

Niemic, C. P., and Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: applying self-determination theory to educational practice. *Theor. Res. Educ.* 7, 133–144. doi: 10.1177/1477878509104318

Pedró, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*. Paris: United Nations Educational, Scientific and Cultural Organization (UNESCO).

Peters, D., Calvo, R. A., and Ryan, R. M. (2018). Designing for motivation, engagement, and wellbeing in digital experience. *Front. Psychol.* 9, 797. doi: 10.3389/fpsyg.2018.00797

Popenici, S. A. D., and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *RPTEL* 12, 22. doi: 10.1186/s41039-017-0062-8

Riva, G., Baños, R. M., Botella, C., Wiederhold, B. K., and Gaggioli, A. (2012). Positive technology: using interactive technologies to promote positive functioning. *Cyberpsychol. Behav. Soc. Network.* 15, 69–77. doi: 10.1089/cyber.2011.0139

Roll, I., and Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *Int. J. Artif. Intellig. Educ.* 26, 582–599. doi: 10.1007/s40593-016-0110-3

Ryan, R. M., and Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY: Guilford Publications.

Schroeder, N. L., Adesope, O. O., and Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *J. Educ. Comput. Res.* 49, 1–39. doi: 10.2190/EC.49.1.a

Steenbergen-Hu, S., and Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K−12 students' mathematical learning. *J. Educ. Psychol.* 105, 970–987. doi: 10.1037/a0032447

Steenbergen-Hu, S., and Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *J. Educ. Psychol.* 106, 331–347. doi: 10.1037/a0034752

Tuomi, I. (2018). *The Impact of Artificial Intelligence on Learning, Teaching, and Education: Policies for the Future,* eds. Y. Punie, R. Vuorikari, and M. Cabrera (Luxembourg: Publications Office of the European Union).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* 46, 197–221. doi: 10.1080/00461520.2011.611369

Verginis, I., Gouli, E., Gogoulou, A., and Grigoriadou, M. (2011). Guiding learners into reengagement through the SCALE environment: an empirical study. *IEEE Trans. Learn. Technol.* 4, 275–290. doi: 10.1109/TLT.2011.20

Winne, P. H., and Baker, R. S. J. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *J. Educ. Data Mining* 5, 1–8. doi: 10.5281/zenodo.3554619

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int. J. Educ. Technol. Higher Educ.* 16, 39. doi: 10.1186/s41239-019-0171-0

frontiers | Frontiers in Education

TYPE Review
PUBLISHED 31 March 2023
DOI 10.3389/feduc.2023.1161777

OPEN ACCESS

EDITED BY
Alona Forkosh Baruch,
Levinsky College of Education, Israel

REVIEWED BY
Tak-Lam Wong,
Douglas College, Canada
Rosanna Yuen-Yan Chan,
The Chinese University of Hong Kong, China

*CORRESPONDENCE
Giuseppe Città
✉ giuseppe.citta@itd.cnr.it

SPECIALTY SECTION
This article was submitted to Digital Learning Innovations, a section of the journal Frontiers in Education

RECEIVED 08 February 2023
ACCEPTED 09 March 2023
PUBLISHED 31 March 2023

CITATION
Gentile M, Città G, Perna S and Allegra M (2023) Do we still need teachers? Navigating the paradigm shift of the teacher's role in the AI era. *Front. Educ.* 8:1161777. doi: 10.3389/feduc.2023.1161777

Do we still need teachers? Navigating the paradigm shift of the teacher's role in the AI era

Manuel Gentile, Giuseppe Città*, Salvatore Perna and Mario Allegra

Institute for Educational Technology, National Research Council of Italy, Palermo, Italy

Through a systematic analysis of the literature, this study analyzes the change in the teacher's role triggered by the integration of AI into educational systems. The picture offered by the systematic analysis of the literature conducted in this study reveals a less than total awareness of the urgency with which the challenges imposed by AI in the educational field must be addressed. We propose a manifesto to guide the evolution of the teachers' role according to the paradigm shift concept proposed by Kuhn in the scientific field.

KEYWORDS

AI and education, paradigm shift, teacher's role, AIED, AI, ChatGPT

1. Introduction

Technological "evolution" has always influenced the world of education by providing new opportunities and challenges for those who form such a foundational part of it as schools and their key players such as teachers and school leaders, students and families.

The new "renaissance" (Tan and Lim, 2018) that AI has been experiencing in recent years, generated by innovations related mainly to deep learning, has stimulated discussion on how advances in AI can influence the educational sector and future educational policies.

In 2018, the EU published a JRC Science for Policy report entitled "The Impact of Artificial Intelligence on Learning, Teaching, and Education" to initiate a well-informed discussion about the state of the art of artificial intelligence (AI) and its potential impact on learning, teaching, and education (Tuomi et al., 2018). Creating a future vision that integrates a careful understanding of our values in education is the key to identifying the contexts in which educational policies could and should intervene.

As indicated in the recent UNESCO report "Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development," the integration of AI in education raises several questions (Pedro et al., 2019).

According to the report "Beijing Consensus on artificial intelligence and education," some of the crucial issues are the need to dynamically review and define teachers' roles and required competencies in the context of teacher policies, strengthen teacher training institutions, and develop appropriate capacity-building programs to prepare teachers to work effectively in AI-rich educational settings.

There is a need to deepen the reflections on the transparency of decision-making processes of AI systems that extends the discussion on ethical issues related to the massive collection of students' data (Miao et al., 2021). Integrating AI techniques into educational processes requires further investigation into the issues of the "digital divide" and social inclusion, the risks associated with such innovations, and the opportunities that technologies offer for handling these issues with new approaches. Above all, a reflection is also needed on

the role of teachers, what skills they should have, and what tools to provide them with to make them conscious actors in these innovation processes.

Starting from the first works that introduced the concept of Artificial Intelligence and Education (AIED) (Cumming et al., 1997; Cumming, 1998; Kay, 2012), several review works have been conducted to offer a systemic view of this phenomenon. The work of Chassignol et al. (2018) analyzes the literature under a four-dimensional framework (content, teaching methods, assessment, and communication) to study the impact of AI in education. The study from Kuka et al. (2022) is a scoping review of where and how AI is used in higher education learning and teaching processes. Another example is the exploratory review of Lameras and Arnab (2022) in which the authors explore the ethical implications of using AI in the educational context, how these technologies can enhance the role of the teacher are discussed, and the applications used and associated teaching and learning practices.

Numerous other reviews on the topic of AI in Education can be found in the literature (e.g., Moreno-Guerrero et al., 2020; Yu and Nazir, 2021; Abdellatif et al., 2022; Dieterle et al., 2022; Megahed et al., 2022, etc.). In addition to the scientific literature, recent books such as Holmes et al. (2019) and Holmes and Porayska-Pomsta (2023) analyze the changes introduced by AI in education.

Of course, these issues are still being also debated in the scientific arena, where a steady increase in studies linking AI and education is evident (Floridi et al., 2018; Holmes and Tuomi, 2022).

Nevertheless, from our point of view what is missing and what prompted us to produce this paper is the need to focus analytically on the paradigm shift in the role of the teacher introduced by the AI era.

Teachers have always been called upon to change their teaching approach by attempting to integrate new technology rather than rejecting it out of hand. However, even at first glance, the potential changes introduced by AI signal a radical change, what can be called a genuine paradigm shift. Therefore, this paper aims to provide a systemic view of this revolution, not by simply offering an overview of the various AI-based tools already available but by trying to grasp the profound changes in the role of the teacher the AI may trigger.

The paper is structured as follows. In Section 2, we discuss the approach used to carry out the review, particularly concerning the choices in defining the search query and coding scheme. Subsequently, in Section 3, we present the results from a general point of view and in detail for each analysis dimension. The paper closes with a discussion that provides a summary of the results and a proposed manifesto to guide the change of the teacher's role in light of the parading shift concept proposed by Kuhn (1962).

2. Materials and methods

The study was done according to the Preferred Reporting Items for Systematic Reviews and Meta-analyzes reporting guideline (PRISMA Checklist) (Page et al., 2021). The search was conducted by accessing the main bibliographic databases such as Web of Science, Scopus to which we have added the ACM Digital Library (DL) to include as much as possible the literature covering computing and information technology.

The research query consists of three main parts. The first lists the terms that allow us to identify the context of artificial intelligence. In particular, we have included both the general terms AI (in contracted and extended form), and the terms machine learning and deep learning, which are often used as synonyms or otherwise identify items of interest for this review. In addition, we have added the extended and contracted form of natural language processing because of the extreme relevance the topic may have in the AI and education field (Litman, 2016). The second group of terms relates to the specific teaching context, where we have included the main aspects of expertise or interest for teachers and teaching processes. Finally, the third group seeks to identify those articles that signal a change, an evolution of the role. Finally, the research focuses on articles published since 2005, a significant date because it identifies the beginning of what has been called the renaissance of AI and coincided with the bursting onto the scene of deep learning.[1]

A clarification is due regarding the absence of the term AIEd (AI in Education, Holmes et al., 2022) among those used within the queries. In the context of our study, we preferred not to limit the analysis to that portion, albeit relevant, of AI that looks explicitly at the educational context identified with the term AIEd (Holmes et al., 2022). Instead, we wanted to analyze the impact of AI in general on the role of the teacher.

In the following box, we report the query used to search the Scopus database. We adapted the query syntax according to the formalism required by each database.

```
TITLE-ABS-KEY
(( "ai" OR "artificial intelligence" OR
 ↪  "deep learning" OR "machine learning"
 ↪  OR "natural language processing" OR
 ↪  "nlp" ) AND
( "educational process" OR "teaching
 ↪  practices" OR "teaching methods" OR
 ↪  "teaching approach" OR "teaching
 ↪  solution" OR "teaching design" OR
 ↪  "teacher development" OR "teaching and
 ↪  learning" OR "pedagogical methods" OR
 ↪  "teacher's role" OR "teachers role" )
 ↪  AND
( transformation OR "new role" OR evolution
 ↪  OR change OR revolution OR enhance OR
 ↪  "paradigm change" OR "paradigm shift"
 ↪  )) AND PUBYEAR > 2005
```

The following inclusion criteria were used for each study: (1) published in English; (2) the study must illustrate, discuss, or theorize the teacher's role in the AI era or must report a study of classroom use if it introduces new AI-based educational practices. Studies that discuss education about AI or studies of articles that discuss only technological and not pedagogical aspects have been excluded.

1 In 2005, the first paper combining the words deep and learn in the title was published (Gomez and Schmidhuber, 2005).

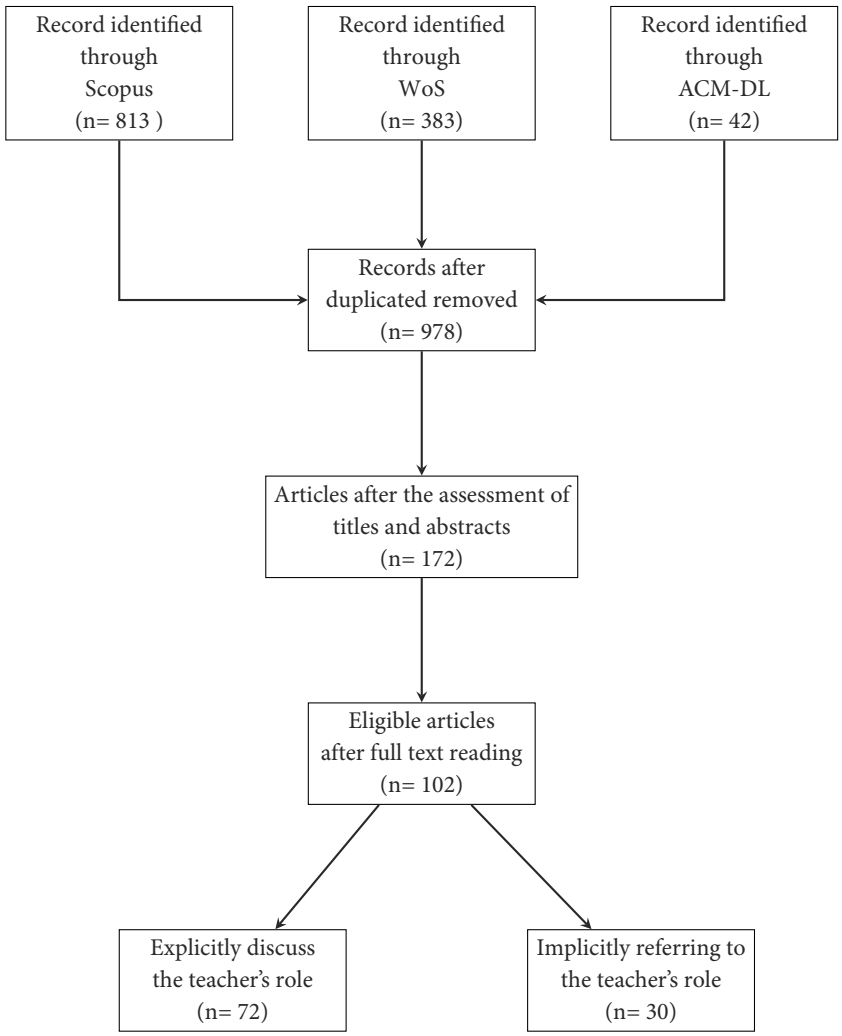


FIGURE 1
Summary of the systematic review process

We searched independently in each database and exported the results in BibTeX format. JabRef software (JabRef Development Team, 2021) was used to merge the three lists and remove duplicates. Subsequently, three reviewers (MG, GC, and SP) independently screened the articles by assessing the articles' titles and abstracts to remove elements inconsistent with the review objective. As shown by Figure 1, this phase led to a drastic reduction in the number of papers (806 papers excluded, amounting to 82.41% of the initial pool). Then the reviewers analyzed the remaining 172 papers by reading the full text and extracting the following information:

1. Year of the publication;
2. Type of the publication;
3. School level;
4. Presence of discussions concerning any of the following topics:

   - Teacher-student interaction
   - Teaching methods and strategies
   - Teaching content
   - Students' assessment and monitoring
   - Teachers' professional development

We identified the topics of investigation (see Figure 2) linking (i) the set of qualities that characterize the teachers' professional development challenges according to Stone Wiske et al. (2001) and (ii) the constituent dimensions of a teaching system identified by Huang et al. (2021).

First, reducing the student-teacher proportion toward the utopian goal of one-to-one interaction has historically been one goal that has driven the implementation of AI systems in the educational sector. In addition to this underlying objective, the first area allows us to explore all the new monitoring and interaction management scenarios AI provides.

Regarding teaching strategies, the innovation challenge of teaching and learning activities connected to integrating new technology into the educational landscape is even more prominent if we consider AI's exponential growth.

AI also impacts the educational content either regarding the need for new professions revolving around the world of AI, as well as considering the possibilities of such techniques to

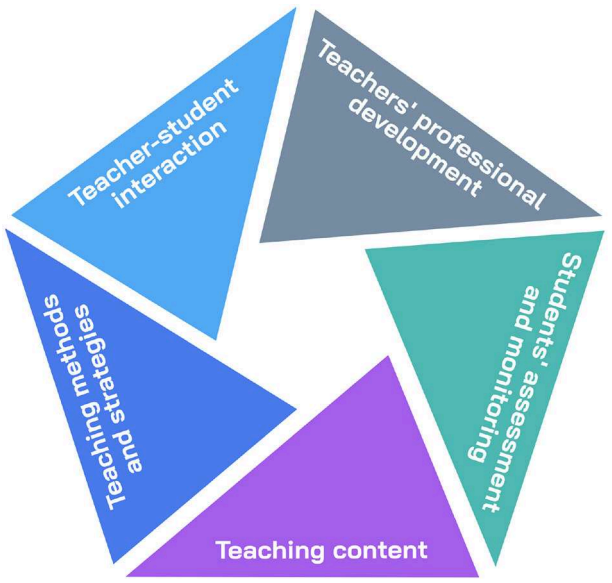


FIGURE 2
The adopted framework of analysis.

support the creation of *ad-hoc* content customized to the needs of learners.

Moreover, the assessment-centered approach is one of the critical goals of professional development in the AI era. AI allows for overcoming the limitations of the traditional summative assessment, making it feasible to move toward continuous formative assessment fully integrated into the teaching process.

To summarize, teachers need professional development opportunities that support them in the transformational process and satisfy these needs.

The reciprocal integration of these dimensions and, consequently, of the elements that are part of each of them makes it possible to look at a teaching system as a whole whose components work coherently. Researching and analysing these dimensions within the papers has meant activating a critical view of the effect of AI on the building blocks of the complex education system.

At the end of the process, 102 papers were considered eligible for review, of which 72 explicitly discuss the teacher's role and 30, while lacking explicit discussion, provide valuable insights for the scope of the review.

3. Results

From the analysis of the selected articles, it emerges that the application of AI in Education is operating and will increasingly operate profound changes on the constitutive pillars of the educational system and on the role the teachers play in it. The trend of annual publications on the study topic shows exponential growth and increasing interest in the scientific community (see Figure 3). The Table 1 shows an even distribution among the dimensions of analysis covered by the articles.

The distribution by country emerging from the analysis of the selected articles shows an interesting result. As highlighted by Figure 4, about half of the articles come from China, sketching a huge gap already with the second country, the United States of America. This gap demonstrates a particular focus of Chinese research concerning the subject matter of this study, namely the analysis of the role of the teacher from the perspective of integrating AI in education.

3.1. Teacher-student interaction

The relationship between teacher and learners, with the interactions that arise and develop within it, is one of the most critical elements in the context of the educational paradigm. The literature analyzed extensively highlights how technological developments, in particular the advent of the artificial intelligence era, have profoundly challenged the teacher-student-teacher interaction models to which we are accustomed.

In the context of the traditional educational paradigm, interactions between teacher and student are predominantly individual, and the moments of feedback and assessment are limited and timed according to the topics covered by the curricula (Liu M. et al., 2022). Within this context, one of the most crucial moments of interaction and feedback is precisely that which takes place simultaneously as the evaluation of the student's performance, which typically follows the presentation of the topics. The classroom setting implicitly forces the teacher to postpone interactions with individual students, but this makes it particularly difficult to assess the effect of his or her teaching and the status of the individual difficulties presented by the various students.

After all, within the traditional educational model, the teacher represents the authority, and the relationship between teacher and learner is hierarchical (Ye, 2021). The transfer of knowledge often occurs in a non-participatory and non-interactive manner and is seen as pouring from a full container into an empty one. Students, inherently characterized by individual needs and peculiarities, are often categorized within families of "similar elements" with the result of standardizing education in a convergent manner to the detriment of its effectiveness. This type of relationship tends not to foster collaboration between teacher and students and, in extreme cases, can lead to negative effects such as absenteeism and dropout (Li, 2021).

The literature shows that, among the elements introduced by the AI era with the most significant impact on the way teachers and learners interact, one of the most important is that of Intelligent Tutoring Systems (ITS). These systems are designed to interact with students and provide them with tutoring intelligently and automatically. They collect data on students' responses and actions to create a model of their knowledge and adapt to their needs. In this way, ITS manage to create a digital profile of the student and provide them with a personal tutor (Chassignol et al., 2018). This feature often leads to the erroneous conclusion that these systems can completely replace the teacher's figure. Although ITS can help enormously in implementing adaptive and personalized teaching and learning strategies, they are not a substitute nor an obstacle in the relationship and interactions between teacher

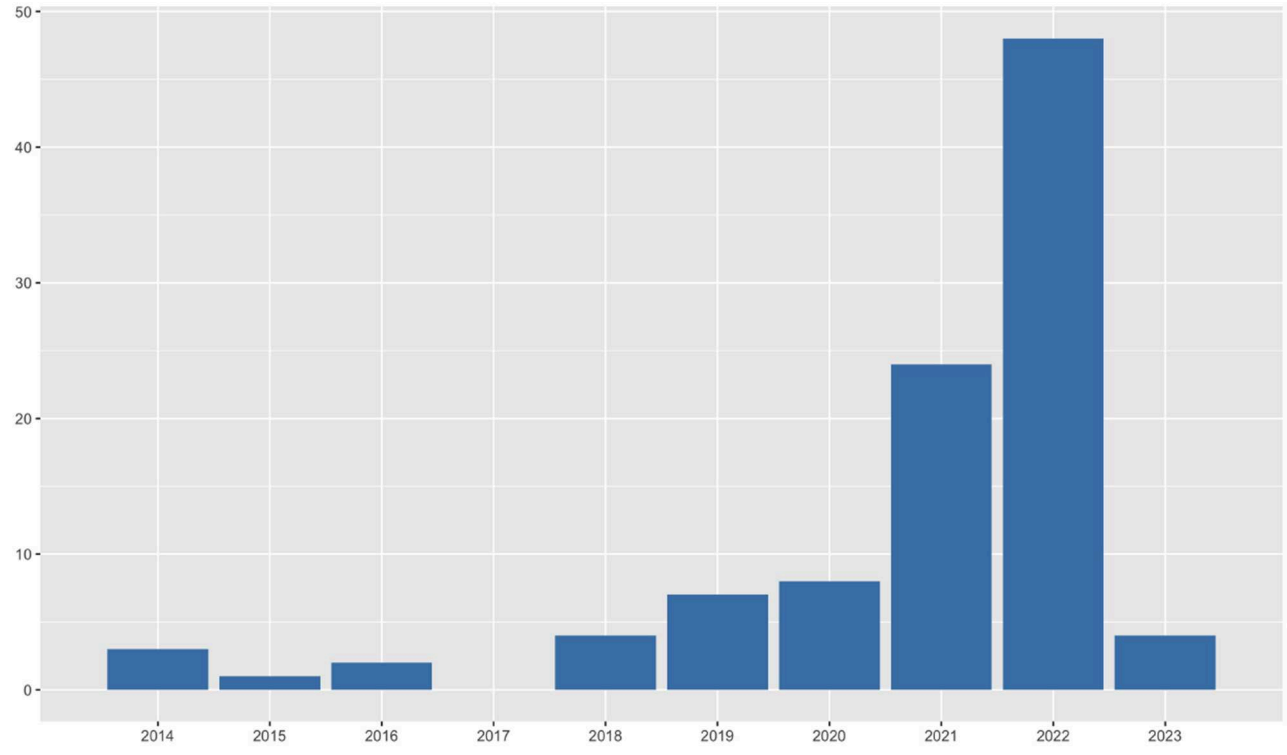


FIGURE 3
Distribution of papers by publication year.

and learners. On the contrary, they can be an essential support tool for the teacher to save valuable time in executing tasks like assessing large numbers of students and presenting teaching materials and resources. Moreover, ITS could provide them with the opportunity to increase the quantity and quality of interactions with students and consequently to identify gaps in learning and teaching at an early stage (Chassignol et al., 2018; Miao and Yao, 2021). Teachers can use the information gathered through ITS to accurately diagnose differences between students and use it to recommend customized and suitable resources (Li, 2021).

Similarly, chatbots and robots can take over most of the interaction related to content and teaching materials and resources (Megahed et al., 2022; Timofeeva and Dorofeeva, 2022; Zhou, 2022). They can automatically answer the most frequent or repetitive questions, enhancing interactions with the teacher and allowing them to become more connected to students' learning strategies and individual needs.

The continuous technological advancements in natural language processing (NLP), which have led to results such as ChatGPT (OpenAI, 2021), only further strengthen the effectiveness of AI-based tools in communicating and interacting with students. Although Artificial Narrow Intelligence (ANI) is making rapid progress (Johri, 2022) and several studies prove the effectiveness of those systems in providing knowledge in various fields (Chassignol et al., 2018; Li, 2021; Su and Yang, 2022), the human teacher remains (to this day) irreplaceable.

Specifically, the human factor is an irreplaceable characteristic of the teacher. The teacher is a guide and reference for students' growth and a compass for their ethical and moral development. In this sense, these tools, which at first glance seem antagonistic to the teacher, are facilitators of the quality interactions that characterize the teaching process.

TABLE 1 Number of papers discussing the coding schema dimensions.

Coding schema dimensions	Number of papers
Teacher-student interaction	37
Teaching methods and strategies	33
Teaching content	27
Students' assessment and monitoring	29
Teachers' professional development	22

Another element explored in the literature that may be complementary to the systems discussed above is the smart classroom (Zhang, 2014; Kowch and Liu, 2018; Zhang Y. et al., 2021; Dimitriadou and Lanitis, 2022, 2023; Liu M. et al., 2022). Although integrating technology within the classroom environment is not a new idea (Rescigno, 1988), constant technological advances in AI have given the idea more and more traction.

A smart or future classroom is intended as a highly integrated environment with sensors and devices capable of automatically controlling and acting on environmental factors (such as temperature and lighting), enhancing communication and real-time interactions both between students and between students

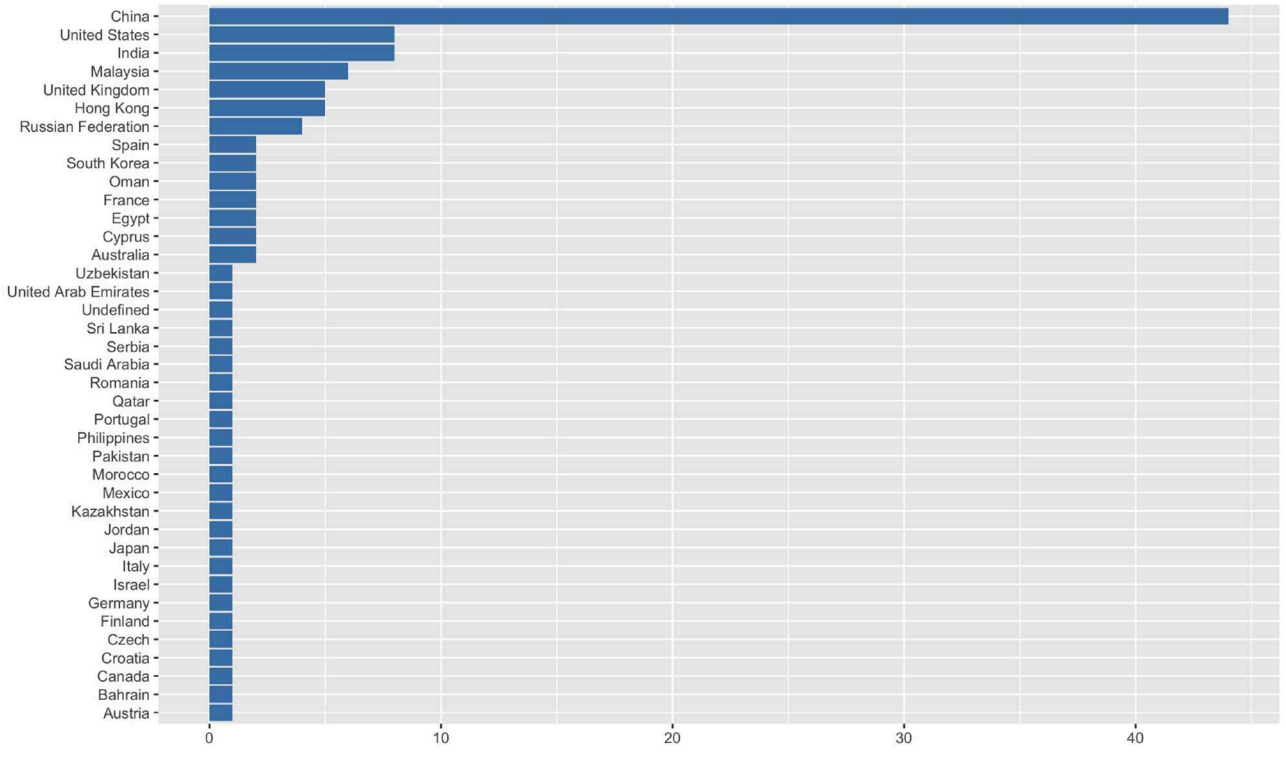


FIGURE 4
Distribution of eligible papers by country.

and teachers (by fostering inter-group collaboration and resource sharing), and overcoming the boundaries and limitations offered by the traditional classroom environment through the Internet and the cloud.

In particular, through the use of Artificial Intelligence of Things (AIoT) and wearable devices, it is possible to capture and monitor students' and teachers' behavioral data in real-time (Zhang Y. et al., 2021; Dimitriadou and Lanitis, 2023). Furthermore, through electronic whiteboards, it is possible to introduce and utilize different types of educational resources that are often difficult to use and present within the traditional classroom context. Another application of smart classroom opportunities is the use of visual feedback (obtained through cameras) to monitor students' attention and emotional state. Through the devices used, it is possible to harness the power of AI to check students' status (both in presence and remotely), obtain real-time information, identify problems related to individual students, and intervene promptly. Despite the countless opportunities and substantial benefits offered by this type of environment, it is nevertheless essential to carefully consider the ethical consequences of such data collection (Dimitriadou and Lanitis, 2022). The learners' (and teachers') right to privacy and security requires a careful evaluation of the data collection, storage and processing protocols. In addition to the costs associated with these types of devices and environments, the delicate nature of data processing and collection can be a strong deterrent for the stakeholders involved. Moreover, the current state of machine learning and AI models does not allow us to exclude potential biases and errors related to false positives (Johri, 2022). This implies that teachers must be able to deeply understand the functioning mechanisms of these systems in order to be able to recognize errors and act in an efficient manner that serves the teaching and pedagogical objectives.

An interesting observation concerning the change in the interactions brought about by AI is that provided by Johri (2022). In his work, the author approaches the topic from a socio-material point of view (Orlikowski, 2002, 2008; Latour, 2007; Suchman and Suchman, 2007; Orlikowski and Scott, 2008; Sørensen, 2009), in which the learning process is strongly dependent on both the social and the material context in which it takes place. According to Johri (2022), the impact of AI on the processes and role of humans within learning practices is fundamentally different from the one technologies had until now. The central element of discontinuity stands in the ability of AI to provide technology with the power and agency to initialize interactions, configuring itself as a communicator on par with the human (Edwards, 2021). The evolution led by AI represents a disruptive change with respect to how we have constructed the concept of agency in the past through social interactions. In this AI era, human-machine communication becomes bi-directional and, above all, can be initiated by both actors. Notifications, alerts and automatic messages are examples of this phenomenon. Technology, which until now has always played a material role, is becoming an agent in its own right. Considering large learning models' rapid innovations and achievements in producing original resources and contents, both textual (OpenAI, 2021) and audiovisual (Ramesh et al., 2021; Singer et al., 2022), it is easy to see how much these communication skills will improve in the future.

In the context of the relationship between teachers and learners, the innovations and disruptions conveyed about by AI are numerous and act at a fundamental level. The primary transformation is about the roles within this relationship. Teachers must learn to become collaborators, mentors and guides of their students (Wang C, 2021; Khusyainov, 2022; Tapalova and Zhiyenbayeva, 2022). Interactions, until now often driven by the concept of authority, must acquire a different character, more driven by symbiotic and increasingly equitable dialogue (Li, 2021), to transform authority into authoritativeness.

AI provides the opportunity to uncover new occasions of interaction, both synchronous and asynchronous, that are freed from the timing and structure of traditional school organization and can focus on those factors that make a person a teacher. These new moments and modes of interaction can be an essential catalyst for implementing proper educational pathways tailored to the needs of individuals. Teachers must cultivate students' passion for learning, ability to think critically, and ability to navigate the sea of information and educational resources surrounding us.

Rather than identifying AI as an antagonist, educators must learn to coexist with it, moving from a binary (student-learner) to a ternary (student-machine-learner) relationship in which interactions are mediated, modified, and sometimes initiated by technology in a way that is enhanced, rather than diminished.

3.2. Teaching methods and strategies

Regarding the methods and strategies implemented by teachers in educational processes, the main contribution of AI relays the shift in the center of gravity of teaching processes and models: from teacher-centered models and processes to learner-centered models and processes (Hou, 2020). This does not mean that the teacher's figure has been given a back seat. It means, instead, that with the advent of AI, this figure, given the current state of the teaching strategies implemented, is invested by profound changes that we will try to explain with the help of the literature analyzed.

The key component that emerges from the examination of the papers is commonly expressed through the expression "personalized learning" (Chassignol et al., 2018). Students' learning rhythms are not standard and the same for all, just as the prior knowledge of each student and the issues related to each course of study are heterogeneous (Ye, 2021). By action of the AI, the personalization of learning paths produces tangible effects on students' learning processes, and this is by an almost revolutionary reconfiguration of the teaching strategies and methods implemented by teachers (Su and Yang, 2022). With the help of tools and techniques such as smart platforms, big data, cloud computing, machine learning, and natural language processing (Litman, 2016; Asgari and Antoniadis, 2021; Liu M. et al., 2022; Liu Y. et al., 2022; Megahed et al., 2022; Dimitriadou and Lanitis, 2023)the teacher can implement comprehensive student monitoring. Thanks to these tools, the teacher is able to collect, represent and analyze in-depth data on students' learning behavior, their learning attitudes and styles, the educational needs of each student and the mutual differences between personal learning paths. Placed in this context, the teacher becomes the actor who, as the first step in an educational pathway, does not introduce knowledge content to the students but possesses the elements to elaborate learner models for each student (Ye, 2021; Yusupova et al., 2022; Dimitriadou and Lanitis, 2023).

The teacher who has this knowledge available should reflect on the strategies and methods used to date and those that are more appropriate today. While in a classical scenario without the contribution of AI, teachers tend to assign the same tasks, lessons and tests to all the students, in the AI era, the teacher can focus on designing innovative ways of teaching. Teachers no longer focus on transmitting homogeneous knowledge contents and designing assessment modalities to verify whether the student has assimilated those specific contents (see Section 3.3) (Liu Y. et al., 2022). In the AI era, the teacher can focus on promoting students' skills like collaboration, autonomy, exploration, problem-solving and creativity. Elements that machines cannot yet emulate (Miao and Yao, 2021). In essence, it is the case that the teacher can be freed from many of the activities that students used to perform and that required his or her help, if not his or her presence: correcting homework, vocabulary training, composing tasks to train numeracy skills, writing, answering frequently asked questions, organizing activities in time. Now, the teacher's work can entirely focus on understanding each student's abilities and level to design targeted teaching paths with the ultimate aim of stimulating the students' personality, self-esteem and potential (Wang C, 2021).

Thanks to the contribution of AI, the teacher of the future will have to focus on the following objectives: cultivating the individual development of students; designing interactive and open teaching even at a distance (in time and space) through the combination of virtual and physical environments; fostering the development of students' autonomous learning, growth and ability to express themselves; considering the flexibility of teaching activities the norm and not the exception; giving particular emphasis in teaching design to the cultivation of the so-called students' 'non-intelligence factors' (the moral character, intelligence, sporting dimension and artistic dimension) (Liu and Wang, 2020; Caijun et al., 2021; Huang and Gupta, 2022). The full realization of the aims mentioned above represents a future of teaching strategies, even if they have already been pursued in numerous educational contexts combined, in some cases, with gamification techniques (e.g., rankings and prizes) and the use of robotic platforms (Vogt et al., 2019) to collect and analyze students' progress and offer the most functional strategy/tool also taking into account any disabilities (Chassignol et al., 2018).

Another essential element that emerged from the literature analysis concerns the extension in time and space of the actions and strategies the teacher can implement thanks to the use of AI. AI allows teachers to characterize their work according to the specific time and space they occurred (before, during and after class) (Yang, 2022). Tasks like the analysis of students' profiles, the design of learning pathways and activities, the organization of assessment sessions and modes in a flexible manner and structural coupling with the environment will assume specific flavors according to the specific timing and place in which these events are to take place. In other words, the teacher has, for the first time, the opportunity to design, develop and implement a systematic teaching model that accompanies the individual learner's learning and personalized

pathways in space and time that goes beyond the classroom environment. Liu Y. et al. (2022) express very clearly this "new" possible scanning of learning paths. The authors highlight how, in a before-class context, learning tasks can be assigned in advance to individual students through a smart platform and how this operation allows the student to explore and study tasks in advance, ask questions or raise doubts. All of these actions can be collected and analyzed by the teacher as feedback data *via* a smart platform and, consequently, used to calibrate individual learning paths to the student's traits and level of knowledge or to modify the design of the entire teaching proposal. Such a range of possible actions clearly also has repercussions on the strategies and methods put into practice in the context of in-class teaching. They are expressed into the possibility for the teacher to design refined teaching: introduction of the topics to be studied in a contextualized manner, active guidance of the student in the exploration of the problems and difficulties recorded in the 'before-class' context, targeted brainstorms on specific topics emerging from the feedback data. The teaching model becomes systematic by implementing specific activities in the 'after-class' context. Within this context, the teacher can pose open questions based on the feedback from the two previous contexts, assign tasks individually based on learning status, and conduct online tutoring sessions to guide each student to summarize and subsume their learning.

An excellent overview of how much AI can impact, and how much it will impact in the future, on teaching strategies and methods is given by Lameras and Arnab (2022). According to them, AI represents a concrete possibility to: (a) support teachers to design adaptive and personalized content and activities appropriate to the knowledge, competence and needs of students, (b) empower teachers and AI agents to collaborate in collecting and analysing student learning and cognitive feedback data, (c) help teachers to step into the shoes of tutors of students' emotional awareness and cultivators of each student's social and affective learning.

3.3. Teaching content

The extent of the changes brought about by the advent of AI is also noteworthy in the production and delivery of teaching content in all educational contexts and at different levels of complexity (Liu and Wang, 2020; Wang C, 2021). The term "teaching content" refers to the body of knowledge and information that teachers teach and that students are expected to learn within a given domain of knowledge (Chassignol et al., 2018). According to Huang et al. (2021), teaching content comprises the knowledge, skills, thoughts and behaviors transmitted by the school to students at all levels.

From the literature analysis, day by day, thanks to the new possibilities offered by AI, new and different types of educational resources and new ways of generating them emerge (Bucea-Manea-oniş et al., 2022; Khusyainov, 2022; Niu, 2022). It is precisely in this direction that "Content Intelligence," a discipline-specific to AI, until now applied to marketing automation, is beginning to operate in the field of education. It opens up the possibility of organizing content and, at the same time, extracting real-time indications on the navigation behavior, fruition and preferences of students in order to implement a customized educational offer.

As emerged in the previous Section 3.2, also in the context of teaching content, personalized learning emerges as the primary effect produced by the action of AI in the educational sphere. This effect disrupts the traditional ways of conceiving, processing and proposing teaching content. Within a conventional didactic approach, teaching contents are the same for all students and static, defined as "closed" because they are hardly modifiable (Ye, 2021; Lameras, 2022). They are organized to be learnt linearly and progressively. According to Hao (2022), the advent of customization makes obsolete the one-way knowledge transfer approach from teachers to learners. It gives way to a focus on the personalized learning processes of students, enabling teachers to organize teaching content that enhances learners' sense of personal fulfillment and helps them learn autonomously. In this perspective, teaching content in general and learning resources in particular change form, structure and how they are generated (Shuguang et al., 2020; Nye et al., 2021).

Thanks to the contribution of the AI, teaching contents move to a new formula in which courses and possible reference texts are accompanied by *ad-hoc* created digital resources and resources from the Internet. These resources can, through machine learning and deep learning, be organized in a multimodal manner and divided into a series of smaller resources that are more manageable and adaptable to the different needs of students. Precisely because of their new multimodal character, they cannot be structured in a monolithic form but must be organized as cross-media and flexible structures so that they can be adapted to the abilities, levels and needs of individual learners. Therefore, contents change starting from their generation process: no longer a static and homogeneous generation but a dynamic and customized one. A generation that makes them dynamic contents of intelligent learning systems that provide personalized paths to the students (Shuguang et al., 2020).

It appears from the analysis of the papers that the trait of personalization of teaching content introduced with the advent of AI means that they acquire new related identifying characteristics. The "new" teaching content will be flexible, manipulatable, explorable, and automatically generated.

They will have flexibility and manipulability as a direct consequence of a personalized teaching approach. The creation of teaching content in an AI context contemplates the need to construct content that can be constantly modified, enhanced, revised and integrated to create original structures perfectly adaptable to the learner's different needs. In this regard, the in-depth examination of the selected papers reveals several examples of the implementation of the flexibility and manipulability features of the teaching content with the help of specific tools and/or teaching strategies. It is crucial that the teaching content be designed in such a way that (a) it can be used by students autonomously through libraries and corpora (e.g., cloud classroom libraries), (b) can be explored collaboratively (Dai, 2021) and continuously (Wang D, 2021) and (c) can be generated and managed automatically according to defined learning objectives (Liu Y. et al., 2022; Schroeder et al., 2022) through specific tools such as automatic question generation tools (Van Campenhout et al., 2021), video content generation tools (Zhang Z. et al., 2021), analytics-based platforms (Conklin, 2016), cloud service solutions or algorithm-based platforms (Alsheref and Fattoh, 2020), e-learning platforms (Khan et al., 2022), natural language processing

and image processing techniques (Sandanayake and Bandara, 2019).

Moreover, the study of the literature reveals that the traits of personalization, flexibility, manipulability, explorability and automatic generation triggered by AI are structural traits of the teaching content of the AI era. For this reason, they are transversal to the single disciplines to which the simple learning contents can be attributed. The analysis of the case studies contained in the analyzed papers reveals a broad distribution across the different fields of knowledge: Anatomy (Abdellatif et al., 2022), Proportional Reasoning (Nye et al., 2021), English and other foreign language teachings (He, 2021; Faustino and Kaur, 2022; Liu and Huang, 2022), Microbial Metabolism (Schroeder et al., 2022), Psychology (Schroeder et al., 2022), Interior Design teaching (Cao and Li, 2022), Social Work Education (Hodgson et al., 2021), Engineering Education (Megahed et al., 2022), Music Design (Dai, 2021), Scientific Writing (Kim and Kim, 2022), Programming Languages (Yusupova et al., 2022), and Translation teaching (Yang, 2022).

3.4. Students' assessment and monitoring

The use of AI in educational assessments is a prominent field of application, and its integration into this process has been extensively studied and discussed in the literature. Assessment is considered a fundamental step in evaluating the impact of AI-powered teaching methods (Luckin et al., 2016). According to the review study conducted by Salas-Pilco et al. (2022), AI and Learning Analytics (LA) techniques have the potential to assist teachers in several activities. Moreover, several studies show that in the teachers' perceptions, AI potentially impacts the evaluation processes. For example, Bucea-Manea-oniş et al. (2022) conducted a study on 139 Romanian and Serbian teachers in HEI, revealing that using AI technologies to assess homework, tests, written assignments, and general student monitoring is an opportunity for them (Bucea-Manea-oniş et al., 2022).

Many of the selected articles in our review refer to assessment as one of the fundamental steps to be considered in analysing new teaching processes guided by AI (Miao and Yao, 2021; Faustino and Kaur, 2022; Lameras, 2022; Lameras and Arnab, 2022; Liu Y. et al., 2022). The level of attention on evaluation is probably due to the close relation to one of the most focused aims of AI applications, namely the personalization of students' learning pathways (Li, 2020; Tapalova and Zhiyenbayeva, 2022). Indeed, personalization can only be thought of with a careful analysis of the student, as stated by Luckin et al. (2016), who identifies the definition of the student model as one of the main issues.

The irruption of AI in assessment processes increases the possibilities regarding the object (what), time (when), and context (where) of evaluation.

Concerning the "what" point, in addition to the level of assessment of knowledge and skills gained by students in particular domains, it is stimulating that some authors focus on analysing students' behaviors and assessing their psychological state. Specifically to the latter point, considering the student's emotional state as an element to be evaluated to facilitate effective learning plays a primary role. According to Huang et al. (2021), AI thus enhances the assessment process by giving more significant importance than before to the assessment of learning processes and individual student development (Lau et al., 2014). A perspective that delineates the potential shift from unidirectional toward bidirectional evaluation (Huang et al., 2021; Hao, 2022).

Liu M. et al. (2022) highlight how the introduction of AI allows teachers to provide prompter evaluation reducing the time delay between the learning process and the feedback to the student ("when"). Consequently, AI potentially enhances the teaching process making the evaluation more pertinent in terms of learning effectiveness and the ability to adapt to students' subjectivity. In other words, AI tools for evaluation enable a shift from summative assessment to adaptive assessment necessary for formative feedback (Lameras, 2022).

Concerning the "where" point, online systems like MOOC or Intelligent-Tutoring Systems represent a natural context in which AI-based evaluation tools could show their potential (Chassignol et al., 2018; Shuguang et al., 2020). Nevertheless, some authors dwell on specific case studies such as assessing students' behavior during a lecture or monitoring students through a video camera as they take, for example, proctored exams (Edwards, 2021; Johri, 2022). As highlighted in the Section 3.1, AI contributes to a shift toward an enhanced school environment outlined by some authors with the terms smart or future classroom (Zhang, 2014; Kowch and Liu, 2018; Zhang Y. et al., 2021; Dimitriadou and Lanitis, 2022; Liu M. et al., 2022).

Of course, introducing video-based monitoring (VbM) for exams and assignments in a classroom environment raises the ethical issue disruptively. Nevertheless, the ethical issue relates to different aspects of the evaluation process, from automatic grading to predictive analysis. As reported by Johri (2022), there are already "systems in use now that help to predict student success based on their prior performance" (Shuguang et al., 2020). In fact, the introduction of AI in assessment processes also opens up new scenarios for analysing prediction scenarios made possible by specific techniques regarding significant aspects such as students' drop-out.

The literature analysis also reveals that connected to the theme of assessment is the theme of teacher support systems for evaluating phenomena such as copying and cheating, often resulting from the use of AI systems themselves, as in the case of translation tools (He, 2021). This issue naturally also reverberates to updating teaching strategies described in the Section 3.2.

Regarding disciplines, case studies related to English Teaching (ET) (Hou, 2020; Li, 2020; Zheng and Zhu, 2021; Liu and Huang, 2022), engineering (Megahed et al., 2022), music (Dai, 2021), anatomy (Abdellatif et al., 2022) and movement monitoring in the field of physical education (Cao et al., 2022) stand out among the selected articles. According to Li (2020), the capability of AI technology to accurately distinguish students' grammar mistakes and the opportunity to strengthen students' abilities to converse thanks to speech recognition features confirm the role of AI as an excellent auxiliary to the work of teachers in ET.

Regarding the school level, most articles introduce the topic of assessment. Nevertheless, of particular interest is the article proposed by Su and Yang (2022), in which a scoping review on the

use of AI in early childhood education is presented. In this review, some papers focus on student assessment and the changes produced by the introduction of AI.

Finally, many of these articles dwell on analysing specific AI techniques, among which emerge neural networks, with particular reference to convolutional networks (CNNs) and Bayesian models, both in classical and deep versions. A cross-cutting aspect of all these papers is the focus on teacher training in this area.

3.5. Teachers' professional development

The picture that emerges from the literature clearly shows how the figure of the teacher today is not sufficiently trained and equipped to deal with the new role that the AI era imposes on them. The most common problem that emerges from the analysis of the papers is precisely the low level (and in some cases the absence) of adequate digital skills (Hou, 2020; Edwards, 2021; Wang C, 2021; Ahmed et al., 2022; Bucea-Manea-oniş et al., 2022; Cao and Li, 2022; Kim and Kim, 2022; Yang, 2022; Dimitriadou and Lanitis, 2023). Whether it stems from the age of the teaching class, or from the habit of using more traditional media and methodologies, today's teachers are not up-to-date with the latest technologies (especially AI-based ones) and rarely have adequate knowledge of the tools available within their subject area. Moreover, in order to cope with the evolution of the educational paradigm introduced by AI, it is necessary for teachers not only to be trained in the use of technologies, with a particular focus on the tools used within their discipline (such as automatic translators in the context of second-language teaching), but also to be instructed in their underlying functioning mechanisms. The mere knowledge of these technologies and tools is in fact less important than the ways in which they are configured, situated and used in teaching practice (Johri, 2022). These tools need to be efficiently integrated within teaching activities, exploiting and assisting the new emerging methodologies and embracing their opportunities within the pedagogical dimension (Liu and Wang, 2020; Nye et al., 2021).

Similarly, just as it is important to train teachers on AI, new technologies, and the mechanisms by which these work, it is also important to structure training courses that will last over time and prevent the educational structure from taking a mere isolated step forward. What is needed is precisely to plan a path whereby continuous training becomes a habit through which teachers can keep up and be ready to interface with an ever-changing market and world in which new generations of students are born, grow and develop natively (He, 2021; Jiang, 2021; Hao, 2022). This problem must be addressed in a systemic manner and cannot be left solely to the teachers. In fact, one of the main reasons why new technologies arrive late in the school context is precisely the tendency to use and introduce only mature technologies (He, 2021), which have been thoroughly proven over the years, but which run the risk of becoming obsolete in a world evolving at the impressive speed at which innovation in AI is traveling.

An interesting result is to be found in the works of Bucea-Manea-oniş et al. (2022) and Kim and Kim (2022): in contrast to the prejudice that sees teachers as opposed to the introduction and use of AI within their own activities, these studies show that in reality educators, especially after direct exposure to the world of artificial intelligence, welcome the change. In fact, the greatest point of fear or insecurity often does not lie in the technology itself, but rather in the ethical and privacy issues surrounding its use within the pedagogical framework (Bucea-Manea-oniş et al., 2022). This suggests that there is within the teaching staff a certain awareness of what was discussed earlier, namely the need to evaluate and learn to integrate these technologies not as mere materials but as pedagogical tools.

For these reasons, and for what has been discussed in Section 3.2, it is important that the training of today's and tomorrow's teachers is systemic and structured to ensure continuous training over time, not limited to the technological part, but complemented by methodological training and the development of emotional, ethical and empathy skills (Miao and Yao, 2021), and that above all this training and these skills are approached holistically (Lameras and Arnab, 2022). Educators of the AI era must be able to be aware of the properties and opportunities offered by technology, they must be able to understand, collect, analyze and interpret the data provided by intelligent systems and integrate this within pedagogical methodologies (both new and old). They must be able to guide students within increasingly personalized educational pathways, and above all to change their role through the creation of ethical relationships with systems and digital assistants, in order to leverage the power of AI to better prepare students for lifelong learning.

4. Conclusions

The review study presented in this article is designed to provide a systematic picture of the critical dimensions related to the teacher figure in which AI plays and will be able to play a role as a catalyst for change. As detailed in the Section 2, starting from the analysis of previous studies that we considered relevant to the purpose of the review, we identified the following dimensions of analysis: teacher-student interaction, teaching methods and strategies, teaching content, students' assessment and monitoring, and teachers' professional development.

As seen in Section 3.1, the literature extensively describes and discusses the current state of the relationship between teachers and learners and the interactions that occur within it. Today, most of these interactions occur individually and in moments that are severely limited both in terms of quantity and timing.

AI-Based tools such as Intelligent Tutoring Systems (ITS), chatbots, and robots are often seen as a threat and an attempt to replace the figure of the teacher, when in fact the literature shows us that they can be important tools through which to create new opportunities for interaction, improving the current state in both quantitative (more interactions) and qualitative (more efficient interactions) terms.

Further new opportunities are provided to us by the development of smart classrooms and new school environments that are highly integrated with technology and AI, forcing us to rethink the ways in which we teach and learn. In particular, one of the most disruptive features of AI is to provide agency to technology, transforming the human-machine relationship from

uni-directional to bi-directional. In the world of today and the future, machines independently and autonomously initiate interactions on a par with humans, and this new way of interacting requires profound ethical and methodological considerations. In light of all this, the teacher must revise their role, learning to coexist with AI and technology, seeing it as a collaborator rather than an antagonist, taking on the figure of an authoritative mentor and guide (especially in the ethical, emotional, and human perspective) and leaving behind that of the knowledge-holding authority.

The shift in the center of gravity of teaching processes and models from teacher-centered to learner-centered through AI frees the teacher from many of those activities that students used to perform and that required his or her direct help or supervision. The central concept of "personalized learning" makes it possible to develop and implement a systematic teaching model that accompanies the individual student's personalized learning pathways in space and time that goes beyond the classroom environment (pre-class, in-class, after-class). The analysis of the literature concerning the teaching methods and strategies has shown that a shift in the center of gravity of the educated process on the student is not, however, matched by a shift in scientific reflection on certain crucial dynamics deeply linked to teaching practices. It emerges that, due to the intervention of AI, teaching methodologies and strategies change, but, although the reflection relating to the wide range of possibilities that can be implemented thanks to AI technology has been expressed in depth, there seems to be a lack of an adequate and systematic reflection on the cognitive implications that these new methodologies entail or will entail. The figure of the teacher, in this context, is not relegated to the background, but the description of the new practices that can be implemented seems to confine it in limbo. Important issues such as the changes that occur in the thinking processes of both the student and the teacher (problem-solving processes, decision-making, critical thinking) as a direct effect of the action of these new methodologies are not examined in depth. The teachers' perception of the novelty is not adequately thematised. It is difficult to deduce from the papers what the skills required of the teacher should be to embrace such a change. And there is no in-depth and exhaustive thematisation of the new skills that the teacher, thanks to the new methods and strategies, will have to cultivate in the student.

Linked to teaching strategies is, of course, the issue of teaching content. The customization of teaching content made possible by AI technology implies that it needs to be, in contrast to the past, flexible, manipulatable, explorable, and automatically generated. This can enable teachers to deal with customisable teaching content that can increase students' sense of personal fulfillment and autonomy. However, this context reveals a clear tendency, perhaps even quite dated but still very risky, to conceive of a new role of the teacher in terms of a non-starring actor. The landscape that emerges about teaching content seems to be built on two main focuses: (1) the student and (2) the AI technology that enables the personalization of resources. Between these two cores would move the teacher who, thanks to the technology he or she has at his or her disposal, can analyze student data to implement the personalization process or, possibly, select, according to teaching objectives, the resources to be included in the courses. There are almost no references to the crucial phases, didactic implications and problems of the customization processes of resources, and what role the teacher can play in them. There is a lack of in-depth focus on the centrality of the teacher in giving methodological direction to the process of constructing teaching content (from design to delivery) and on the skills that the teacher must acquire in the management of AI technology both in the phases of the teaching content generation process and in the delivery phases.

The literature review shows that evaluation is one of the most debated topics when considering the application of AI in educational processes. Firstly, assessment has intrinsic value as it is considered a crucial step in teaching methods. Furthermore, AI-based assessment exploits models and techniques in which AI has proven to be particularly effective, such as modeling and classification tasks. The analysis reported in Section 3.4, suggests that AI and learning analytics can help teachers in various activities and positively influence teaching processes. Furthermore, the evaluation also plays an essential role in other processes related to the application of AI in education, such as the personalization of student learning paths. AI integration in assessment processes could enhance it by extending the what, when, and where (in what context) to evaluate the student. Several studies suggest how AI can support the teacher during the assessment process by fostering a greater focus on learning processes and individual student development and enabling faster assessments and formative feedback.

With regard to the professional level and competency framework of teachers, as discussed in Section 3.5, the literature clearly emphasizes that substantial change is needed in order to cope with the evolutionary wave that the AI era brings. The teacher of tomorrow needs careful training that will enable them not only to acquire the necessary digital competences and skills but, even more importantly, to deeply understand the underlying mechanisms of how these new technologies work, so as to be able to integrate and situate them within the didactic pathways in a way that serves pedagogical purposes. It is necessary for teacher training to move from the sphere of pure knowledge of the relevant subject to that of the higher-level cognitive processes that affect learning, so as to be able to make the necessary change of role and truly prepare students for a personalized lifelong learning path. The teacher of the AI era must be a charismatic, empathetic educator able to build ethical relationships and interactions with the intelligences and digital tools that will assist them in their work. Moreover, it is important that this training embraces all these elements in a holistic manner, and above all that it is systematized at an organizational level so as to create continuous training paths that keep teachers up-to-date and ready to face tomorrow's developments.

The picture offered by the systematic analysis of the literature conducted in this study reveals a less than total awareness of the urgency with which the challenges imposed by AI in the educational field must be addressed. For this reason, we propose a kind of manifesto (see Figure 5) for guiding the change of the teacher's role that can reaffirm a "new centrality" of the role, forcefully countering the idea that it can be relegated to a mere mediator or tutor of a path built by "an artificial intelligence." As described by Johri (2022), this urgency originates from the enormous difference introduced by AI compared to other technologies from the point of view of agency. The autonomy that characterizes such technologies, their ability to be initiators of interaction with students, and the complexity of the

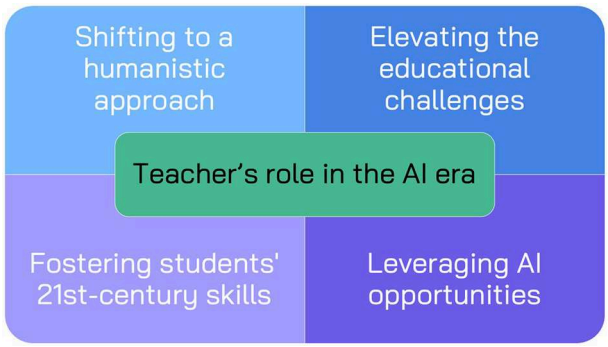


FIGURE 5
A manifesto for the new teacher's role in the AI era.

tasks that AI can already perform and increasingly will be able to do, imposes an evolution of the teacher's role. An evolution that can preserve, or perhaps restore, that beneficial authoritativeness that makes the teacher the point of reference in the student's growth path.

Such a manifesto must, in our opinion, start from a few main points:

- Shifting the teaching objectives from a disciplinary to a "humanistic" approach by focusing on the individual as a person and as a social group member. The teacher should play a more significant role in shaping people, their brains, souls, and moral values than before.
- Elevating the level of the challenges posed to our students. In the AI era, the teacher can no longer ask students the same outcomes that they are used to asking in the past. We need to demand a quantum leap toward students able to actively learn, discover problems, communicate and interact, and deal with complex problems.
- Fostering the development of students' twenty-first-century skills. Teachers should focus toward social skills like collaboration, autonomy and exploration as well as the high-level cognitive processes that characterize them (e.g., critical thinking, problem-solving, etc.).
- Leveraging the opportunities AI provides for designing and implementing innovative teaching methods, managing workload, and extending and enhancing the educational space-time continuum.

Fostering this paradigm shift cannot work only through groundwork on the technological skills of the teacher. Promoting the teachers' awareness about the points listed in our manifesto is a must to do action for all the national educational systems. Teachers should be conscious of the need to become the principal actor of a continuous innovation process from methodological, psychological and cognitive points of view.

We like to conclude this paper with a paraphrase of Kuhn's statement about the paradigm shift in science. To do so, we have taken the liberty of substituting the term "scientist" with the term "teacher" and the term "research" with the term "teaching" in an excerpt from the tenth chapter of the book "The Structure of Scientific Revolutions" (Kuhn, 1962), which is entitled "Revolutions as Changes in Worldview."

The portion of the following text is a perfect synopsis of what we have attempted to depict in this paper.

> *Nevertheless, paradigm changes do cause* ***teachers*** *[scientists] to see the world of their* ***teaching*** *[research] differently. In so far as their only recourse to that world is through what they see and do, we may want to say that after a revolution* ***teachers*** *[scientists] are responding to a different world. [...] Therefore, at times of revolution, when the* ***traditional educational methods*** *[normal-scientific tradition] changes,* ***teachers'*** *[scientists'] perception of his environment must be re-educated - in some familiar situations he must learn to see a new gestalt. [It happens that at the beginning of this process of change, the teacher has to] puts on goggles fitted with inverting lenses and initially sees the entire world upside down. At the start, his perceptual apparatus functions as it had been trained to function in absence of the goggles, and the result is extreme disorientation, an acute personal crisis. But after the [teacher] has begun to learn to deal with his new world, his entire visual field flips over, usually after an intervening period in which vision is simply confused. [...] Literally, as well as metaphorically, the man accustomed to inverting lenses has undergone a revolutionary transformation of vision.*

Author contributions

MG: conceptualization, data curation, formal analysis, methodology, supervision, visualization, writing—original draft, writing—review, and editing. GC and SP: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft, writing—review, and editing. MA: conceptualization, methodology, supervision, visualization, writing—review, and editing. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abdellatif, H., Al Mushaiqri, M., Albalushi, H., Al-Zaabi, A. A., Roychoudhury, S., and Das, S. (2022). Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. *Int. J. Environ. Res. Public Health* 19, 14209. doi: 10.3390/ijerph192114209

Ahmed, S., Khalil, M. I., Chowdhury, B., Haque, R., bin, S., Senathirajah, A. R., et al. (2022). Motivators and barriers of artificial intelligent (AI) based teaching. *Eurasian J. Educ. Res*. 100, 74–89. doi: 10.14689/ejer.2022.100.006

Alsheref, F. K., and Fattoh, I. E. (2020). "Medical text annotation tool based on IBM Watson platform," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Coimbatore: IEEE), 1312–1316.

Asgari, H., and Antoniadis, G. (2021). "Mobile artefacts and language teaching, the example of the spoc+ platform," in *17th International Conferences Mobile Learning 2021 (ML 2021)* (Lisbonne).

Bucea-Manea-oni,ş, R., Kuleto, V., Gudei, S. C. D., Lianu, C., Lianu, C., Ili,ć, M. P., et al. (2022). Artificial intelligence potential in higher education institutions enhanced learning environment in romania and serbia. *Sustainability* 14, 5842. doi: 10.3390/su14105842

Caijun, W., Xi, J., and Zhenzhou, Z. (2021). "Analysis of systematic reform of future teaching in the age of artificial intelligence," in *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)* (Dali), 704–707. doi: 10.1109/ICAIE53562.2021.00154

Cao, F., Xiang, M., Chen, K., and Lei, M. (2022). Intelligent physical education teaching tracking system based on multimedia data analysis and artificial intelligence. *Mobile Inf. Syst*. 2022, 1–11. doi: 10.1155/2022/7666615

Cao, H., and Li, H. (2022). "Research and innovation of interior design teaching method based on artificial intelligence technology "promoting teaching with competition,"" in *Frontier Computing*, eds J. C. Hung, N. Y. Yen, and J.-W. Chang (Singapore: Springer Nature Singapore), 780–789.

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. (2018). Artificial intelligence trends in education: a narrative overview. *Procedia Comput. Sci*. 136, 16–24. doi: 10.1016/j.procs.2018.08.233

Conklin, T. A. (2016). *Knewton An Adaptive Learning Platform*. Available online at: https://www.knewton.com/

Cumming, G. (1998). Artificial intelligence in education: an exploration. *J. Comput. Assist. Learn*. 14, 251–259. doi: 10.1046/j.1365-2729.1998.1440251.x

Cumming, G., Sussex, R., Cropp, S., and McDougall, A. (1997). "Learner modelling: lessons from expert human teachers," in *Artificial Intelligence in Education: Knowledge and Media in Learning Systems, Volume 39 of Frontiers in Artificial Intelligence and Applications, 8th World Conference on Artificial Intelligence in Education - Knowledge and Media in Learning Systems (AI-ED 97)*, eds B. duBoulay and R. Mizoguchi (Kobe), 577–579.

Dai, D. D. (2021). Artificial intelligence technology assisted music teaching design. *Sci. Program*. 2021, 1–10. doi: 10.1155/2021/9141339

Dieterle, E., Dede, C., and Walker, M. (2022). The cyclical ethical effects of using artificial intelligence in education. *AI Soc*. 2022, 1–11. doi: 10.1007/s00146-022-01497-w

Dimitriadou, E., and Lanitis, A. (2022). "The role of artificial intelligence in smart classes: a survey," in *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)* (Palermo: IEEE), 642–647.

Dimitriadou, E., and Lanitis, A. (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learn. Environ*. 10, 12. doi: 10.1186/s40561-023-00231-3

Edwards, B. I. (2021). *Emerging Trends in Education: Envisioning Future Learning Spaces and Classroom Interaction*. Singapore: Springer Singapore.

Faustino, A., and Kaur, I. (2022). Artificial intelligence and machine learning: future of education. *AIP Conf. Proc*. 2555, 050031. doi: 10.1063/5.0109332

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4people— ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 28, 689–707. doi: 10.1007/s11023-018-9482-5

Gomez, F. J., and Schmidhuber, J. (2005). "Co-evolving recurrent neurons learn deep memory pomdps," in *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, GECCO '05* (New York, NY: Association for Computing Machinery), 491–498.

Hao, X. (2022). Innovation in teaching method using visual communication under the background of big data and artificial intelligence. *Mobile Inf. Syst*. 2022, 1–9. doi: 10.1155/2022/7315880

He, Y. (2021). Challenges and countermeasures of translation teaching in the era of artificial intelligence. *J. Phys. Conf. Ser*. 1881, 022086. doi: 10.1088/1742-6596/1881/2/022086

Hodgson, D., Goldingay, S., Boddy, J., Nipperess, S., and Watts, L. (2021). Problematising artificial intelligence in social work education: challenges, issues and possibilities. *Br. J. Soc. Work* 52, 1878–1895. doi: 10.1093/bjsw/bcab168

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*, Boston, MA: Center for Curriculum Redesign.

Holmes, W., and Porayska-Pomsta, K. (2023). *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*. New York, NY: Routledge; Taylor & Francis Group.

Holmes, W., and Tuomi, I. (2022). State of the art and practice in scpAI/scp in education. *Eur. J. Educ*. 57, 542–570. doi: 10.1111/ejed.12533

Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., and Dimitrova, V. (2022). *Artificial Intelligence and Education: A Critical View Through the Lens of Human rights, Democracy and the Rule of Law*. Council of Europe.

Hou, Y. (2020). "Foreign language education in the era of artificial intelligence," in *Big Data Analytics for Cyber-Physical System in Smart City*, eds M. Atiquzzaman, N. Yen and Z. Xu (Singapore: Springer Singapore), 937–944.

Huang, J., Shen, G., and Ren, X. (2021). Connotation analysis and paradigm shift of teaching design under artificial intelligence technology. *Int. J. Emerg. Technol. Learn*. 16, 73–86. doi: 10.3991/ijet.v16i05.20287

Huang, L., and Gupta, P. (2022). An empirical study of integrating information technology in english teaching in artificial intelligence era. *Sci. Program*. 2022, 5097. doi: 10.1155/2022/6775097

JabRef Development Team (2021). *Jabref – An Open-Source, Cross-Platform Citation and Reference Management Software, Version 5.1*. Available online at: https://www.jabref.org

Jiang, Z. (2021). "Discussion on artificial intelligence and information technology application of in the teaching of ideological and political courses," in *2021 4th International Conference on Information Systems and Computer Aided Education, ICISCAE 2021* (New York, NY: Association for Computing Machinery), 881–884. doi: 10.1145/3482632.3483042

Johri, A. (2022). Augmented sociomateriality: implications of artificial intelligence for the field of learning technology. *Res. Learn. Technol*. 30, 2642. doi: 10.25304/rlt.v30.2642

Kay, J. (2012). AI and education: grand challenges. *IEEE Intell. Syst*. 27, 66–69. doi: 10.1109/MIS.2012.92

Khan, A., Hasana, M. K., Ghazal, T. M., Islam, S., Alzoubi, H. M., Mokhtar, U. A., et al. (2022). "Collaborative learning assessment *via* information and communication technology," in *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)* (Ho Chi Minh City: IEEE).

Khusyainov, T. M. (2022). *Uberization of Education: Critical Analysis*. Cham: Springer International Publishing.

Kim, N. J., and Kim, M. K. (2022). Teacher's perceptions of using an artificial intelligence-based educational tool for scientific writing. *Front. Educ*. 7, 755914. doi: 10.3389/feduc.2022.755914

Kowch, E. G., and Liu, J. C. (2018). "Principles for teaching, leading, and participatory learning with a new participant: AI," in *2018 International Joint Conference on Information, Media and Engineering (ICIME)* (Osaka), 320–325. doi: 10.1109/ICIME.2018.00075

Kuhn, T. (1962). *The Structure of Scientific Revolutions. International Encyclopedia of Unified Science: Foundations of the unity of Science v. 2*, Chicago: University of Chicago Press.

Kuka, L., Hörmann C., and Sabitzer, B. (2022). *Teaching and Learning with AI in Higher Education: A Scoping Review*. Cham: Springer International Publishing.

Lameras, P. (2022). "A vision of teaching and learning with AI," in *2022 IEEE Global Engineering Education Conference (EDUCON)* (Tunis: IEEE), 1796–1803.

Lameras, P., and Arnab, S. (2022). Power to the teachers: an exploratory review on artificial intelligence in education. *Information* 13, 14. doi: 10.3390/Info13010014

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Lau, T. P., Wang, S., Man, Y., Yuen, C. F., and King, I. (2014). "Language technologies for enhancement of teaching and learning in writing," in *Proceedings of the 23rd International Conference on World Wide Web* (ACM).

Li, J. (2021). "Research on intimacy between teachers and students in english classrooms in the context of artificial intelligence," in *2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era* (Nicosia: FoNeS-AIoT), 43–46. doi: 10.1109/FoNeS-AIoT54873.2021.00019

Li, Y. (2020). "Application of artificial intelligence in higher vocational english teaching in the information environment," in *Innovative Computing*, eds C.-T. Yang, Y. Pei, and J.-W. Chang (Singapore: Springer Singapore), 1169–1173.

Litman, D. (2016). Natural language processing for enhancing teaching and learning. *Proc. AAAI Conf. Artif. Intell.* 30, 9879. doi: 10.1609/aaai.v30i1.9879

Liu, J., and Wang, S. (2020). "The change of teachers role in teaching under the environment of "artificial intelligence +."" in *2020 International Conference on Artificial Intelligence and Education (ICAIE)* (Tianjin, China), 98–102. doi: 10.1109/ICAIE50891.2020.00030

Liu, M., Zhou, R., Dai, J., Feng, X., and Tang, Y. (2022). Analysis and practice of using modern information technology for classroom teaching mode reform. *Mob. Inf. Syst*. 2022, 2565735. doi: 10.1155/2022/2565735

Liu, X., and Huang, X. (2022). Design of artificial intelligence-based english network teaching (AI-ENT) system. *Math. Probl. Eng*. 2022, 1–12. doi: 10.1155/2022/1849430

Liu, Y., Chen, L., and Yao, Z. (2022). The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes. *Front. Psychol*. 13, 929175. doi: 10.3389/fpsyg.2022.929175

Luckin, R., Holmes, W., Griffiths, M., and Forcier, L. B. (2016). *Intelligence unleashed: An argument for ai in education*. Technical report, London.

Megahed, N. A., Abdel-Kader, R. F., and Soliman, H. Y. (2022). "Post-pandemic education strategy: framework for artificial intelligence-empowered education in engineering (aied-eng) for lifelong learning," in *The 8th International Conference on Advanced Machine Learning and Technologies and Applications (AMLTA2022)*, eds A. E. Hassanien, R. Y. Rizk, V. Snášel, and R. F. Abdel-Kader (Cham: Springer International Publishing), 544–556.

Miao, F., Holmes, W., Huang, R., Zhang, H., et al. (2021). *AI and Education: A Guidance for Policymakers*. UNESCO Publishing.

Miao, Y., and Yao, Y. (2021). "Professional development of college teachers in the era of artificial intelligence: role rebuilding and development path," in *Application of Intelligent Systems in Multi-modal Information Analytics*, eds V. Sugumaran, Z. Xu, and H. Zhou (Cham: Springer International Publishing), 618–626. doi: 10.1007/978-3-030-51431-0_89

Moreno-Guerrero, A.-J., López-Belmonte, J., Marín-Marín, J.-A., and Soler-Costa, R. (2020). Scientific development of educational artificial intelligence in web of science. *Future Internet* 12, 124. doi: 10.3390/fi12080124

Niu, L. (2022). Analysis of multimodal teaching of college english under the background of artificial intelligence. *Security Commun. Netw*. 2022, 3833106. doi: 10.1155/2022/3833106

Nye, B. D., Shiel, A., Olmez, I. B., Mittal, A., Latta, J., Auerbach, D., et al. (2021). Virtual agents for real teachers: applying AI to support professional development of proportional reasoning. *Int. Flairs Conf. Proc*. 34, 128574. doi: 10.32473/flairs.v34i1.128574

Open,AI. (2021). *Gpt-3: The third-generation generative pretrained transformer language model*. Technical report.

Orlikowski, W. J. (2008). "Using technology and constituting structures: a practice lens for studying technology in organizations," in *Resources, Co-Evolution and Artifacts. Computer Supported Cooperative Work* (London: Springer). doi: 10.1007/978-1-84628-901-9_10

Orlikowski, W. J. (2002). Special issue: Knowledge, knowing, and organizations: knowing in practice: enacting a collective capability in distributed organizing. *Organ. Sci*. 13, 249–273. doi: 10.1287/orsc.13.3.249.2776

Orlikowski, W. J., and Scott, S. V. (2008). 10 sociomateriality: challenging the separation of technology, work and organization. *Acad. Manag. Ann*. 2, 433–474. doi: 10.5465/19416520802211644

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372, n160. doi: 10.1136/bmj.n160

Pedro, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development*. Technical report.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., et al. (2021). Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092. doi: 10.48550/arXiv.2102.12092

Rescigno, R. C. (1988). Practical implementation of educational technology. The GTE/GTEL smart-classroom. The hueneme school district experience. *Acad. Achiev*. 10, 1–27.

Salas-Pilco, S. Z., Xiao, K., and Hu, X. (2022). Artificial intelligence and learning analytics in teacher education: a systematic review. *Educ. Sci*. 12, 569. doi: 10.3390/educsci12080569

Sandanayake, T. C., and Bandara, A. M. (2019). Automated classroom lecture note generation using natural language processing and image processing techniques. *Int. J. Adv. Trends Comput. Sci. Eng*. 8, 1920–1926. doi: 10.30534/ijatcse/2019/16852019

Schroeder, K. T., Hubertz, M., Van Campenhout, R., and Johnson, B. G. (2022). Teaching and learning with ai-generated courseware: lessons from the classroom. *Online Learn*. 26, 73–87. doi: 10.24059/olj.v26i3.3370

Shuguang, L., Zheng, L., and Lin, B. (2020). "Impact of artificial intelligence 2.0 on teaching and learning," in *Proceedings of the 2020 9th International Conference on Educational and Information Technology, ICEIT 2020* (New York, NY: Association for Computing Machinery), 128–133.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., et al. (2022). Make-a-video: text-to-video generation without text-video data. *ArXiv*, abs/2209.14792. doi: 10.48550/arXiv.2209.14792

Sørensen, E. (2009). "The materiality of learning," in *Learning in Doing: Social, Cognitive and Computational Perspectives* (Cambridge: Cambridge University Press), 177–194.

Stone Wiske, M., Sick, M., and Wirsig, S. (2001). New technologies to support teaching for understanding. *Int. J. Educ. Res*. 35, 483–501. doi: 10.1016/S0883-0355(02)00005-8

Su, J., and Yang, W. (2022). Artificial intelligence in early childhood education: a scoping review. *Comput. Educ. Artif. Intell*. 3, 100049. doi: 10.1016/j.caeai.2022.100049

Suchman, L., and Suchman, L. A. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press.

Tan, K.-H., and Lim, B. P. (2018). The artificial intelligence renaissance: deep learning and the road to human-level machine intelligence. *APSIPA Trans. Signal Inf. Process. 7*, E6. doi: 10.1017/ATSIP.2018.6

Tapalova, O., and Zhiyenbayeva, N. (2022). Artificial intelligence in education: aied for personalised learning pathways. *Electron. J. e-Learn*. 20, 639–653. doi: 10.34190/ejel.20.5.2597

Timofeeva, E. G., and Dorofeeva, A. A. (2022). Digital transformation of the russian historical education. *Galactica Media J. Media Stud*. 4, 284–294. doi: 10.46539/gmd.v4i4.350

Tuomi, I., Cabrera Giraldez, M., Vuorikari, R., and Punie, Y. (2018). "The impact of artificial intelligence on learning, teaching, and education," in *Anticipation and foresight KJ-NA-29442-EN-N (online)* (Luxembourg: KJ-NA-29442-EN-E).

Van Campenhout, R., Dittel, J. S., Jerome, B., and Johnson, B. G. (2021). "Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation," in *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education* (CEUR Workshop Proceedings). Available online at: http://ceur-ws.org/Vol-2895/paper06.pdf

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). "Second language tutoring using social robots: a large-scale study," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Daegu: IEEE).

Wang, C. (2021). "The value orientation of teachers' role in artificial intelligence teaching environment based on information technology," in 2021 *4th International Conference on Information Systems and Computer Aided Education, ICISCAE 2021* (New York, NY: Association for Computing Machinery), 951–954.

Wang, D. (2021). "Changes and challenges: a study on the application of artificial intelligence technology in college english teaching," in *2021 4th International Conference on Information Systems and Computer Aided Education* (New York, NY: ACM), 1–6. doi: 10.1145/3482632.3483151

Yang, C. (2022). "The application of artificial intelligence in translation teaching," in *Proceedings of the 4th International Conference on Intelligent Science and Technology (ICIST)* (New York, NY: ACM), 1–7. doi: 10.1145/3568923.3568933

Ye, Z. Q. (2021). "Dual logic of teacher role transformation based on artificial intelligence," in *2021 2nd International Conference on Big Data and Informatization Education (ICBDIE)* (Hangzhou), 282–286. doi: 10.1109/ICBDIE52740.2021.00070

Yu, H., and Nazir, S. (2021). Role of 5g and artificial intelligence for research and transformation of english situational teaching in higher studies. *Mobile Inf. Syst*. 2021, 3414. doi: 10.1155/2021/3773414

Yusupova, S. B., Sultanov, O. R., Baltayev, R. S., and Bekchanov, F. A. (2022). "The advantage of using e-learning in teaching students programming languages," in *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)* (Yekaterinburg: IEEE).

Zhang, J. (2014). "Reconstructing new space for teaching and learning: the future classroom," in *Hybrid Learning. Theory and Practice*, eds S. K. S. Cheung, J. Fong, J. Zhang, R. Kwan, and L. F. Kwok (Cham: Springer International Publishing), 49–55.

Zhang, Y., Ning, Y., Li, B., and Liu, Y. (2021). "An innovative classroom teaching technology assisted by artificial intelligence of things," in *2021 2nd International Conference on Information Science and Education (ICISE-IE)* (Chongqing: IEEE), 1661–1664.

Zhang, Z., Hu, W., and Yang, Z. (2021). "Research on the innovation and development of visual communication design in the new media era," in *Advances in Social Science, Education and Humanities Research* (Paris: Atlantis Press).

Zheng, S., and Zhu, S. (2021). "A study of college english translation teaching in the age of artificial intelligence," in *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, 998–1000.

Zhou, Y. (2022). Research on innovative strategies of college students' english teaching under the background of artificial intelligence. *Appl. Math. Nonlinear Sci*. 2022, 272. doi: 10.2478/amns.2021.2.00272

frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 05 May 2023
DOI 10.3389/frai.2023.1151391

OPEN ACCESS

EDITED BY
Manuel Gentile,
Institute for Educational Technology-National Research Council of Italy, Italy

REVIEWED BY
Heinrich Söbke,
Bauhaus-Universität Weimar, Germany
Chiara Panciroli,
University of Bologna, Italy

*CORRESPONDENCE
Sruti Mallik
✉ sruti.mallik@wustl.edu

†These authors have contributed equally to this work and share first authorship

RECEIVED 27 January 2023
ACCEPTED 06 April 2023
PUBLISHED 05 May 2023

CITATION
Mallik S and Gangopadhyay A (2023) Proactive and reactive engagement of artificial intelligence methods for education: a review. *Front. Artif. Intell.* 6:1151391. doi: 10.3389/frai.2023.1151391

Proactive and reactive engagement of artificial intelligence methods for education: a review

Sruti Mallik*† and Ahana Gangopadhyay†

Washington University in St. Louis, St. Louis, MO, United States

The education sector has benefited enormously through integrating digital technology driven tools and platforms. In recent years, artificial intelligence based methods are being considered as the next generation of technology that can enhance the experience of education for students, teachers, and administrative staff alike. The concurrent boom of necessary infrastructure, digitized data and general social awareness has propelled these efforts further. In this review article, we investigate how artificial intelligence, machine learning, and deep learning methods are being utilized to support the education process. We do this through the lens of a novel categorization approach. We consider the involvement of AI-driven methods in the education process in its entirety—from students admissions, course scheduling, and content generation in the *proactive* planning phase to knowledge delivery, performance assessment, and outcome prediction in the *reactive* execution phase. We outline and analyze the major research directions under proactive and reactive engagement of AI in education using a representative group of 195 original research articles published in the past two decades, i.e., 2003–2022. We discuss the paradigm shifts in the solution approaches proposed, particularly with respect to the choice of data and algorithms used over this time. We further discuss how the COVID-19 pandemic influenced this field of active development and the existing infrastructural challenges and ethical concerns pertaining to global adoption of artificial intelligence for education.

KEYWORDS

artificial intelligence applications (AIA), artificial intelligence for education (AIEd), technology enhanced learning, machine learning, artificial intelligence for social good (AI4SG)

1. Introduction

Integrating computer-based technology and digital learning tools can enhance the learning experience for students and knowledge delivery process for educators (Lin et al., 2017; Mei et al., 2019). It can also help accelerate administrative tasks related to education (Ahmad et al., 2020). Therefore, researchers have continued to push the boundaries of including computer-based applications in classroom and virtual learning environments. Specifically in the past two decades, artificial intelligence (AI) based learning tools and technologies have received significant attention in this regard. In 2015, the United Nations General Assembly recognized the need to impart quality education at primary, secondary, technical, and vocational levels as one of their seventeen sustainable development goals or SDGs (United Nations, 2015). With this recognition, it is anticipated that research and development along the frontiers of including artificial intelligence for education will continue to be in the spotlight globally (Vincent-Lancrin and van der Vlies, 2020).

In the past there has been considerable discourse about how adoption of AI-driven methods for education might alter the course of how we perceive education (Dreyfus, 1999; Feenberg, 2017). However, in many of the earlier debates, the full potential of artificial intelligence was not recognized due to lack of supporting infrastructure. It was not until very recently that AI-powered techniques could be used in classroom environments. Since the beginning of the twenty-first century, there has been a rapid progress in the semiconductor industry in manufacturing chips that can handle computations at scale efficiently. In fact, in the coming decade too it is anticipated that this growth trajectory will continue with focus on wireless communication, data storage and computational resource development (Burkacky et al., 2022). With this parallel ongoing progress, using AI-driven platforms and tools to support students, educators, and policy-makers in education appears to be more feasible than ever.

The process of educating a student begins much before the student starts attending lectures and parsing lecture materials. In a traditional classroom education setup, administrative staff, and educators begin preparations related to making admissions decisions, scheduling of classes to optimize resources, curating course contents, and preliminary assignment materials several weeks prior to the term start date. In an online learning environment, similar levels of effort are put into structuring the course content and marketing the course availability to students. Once the term starts, the focus of educators is to deliver the course material, give out and grade assignments to assess progress and provide additional support to students who might benefit from that. The role of the students is to regularly acquire knowledge, ask clarifying questions and seek help to master the material. The role of administrative staff in this phase is less hands-on—they remain involved to ensure smooth and efficient overall progress. It is therefore a multi-step process involving many inter-dependencies and different stakeholders. Throughout this manuscript we refer to this multi-step process as the *end-to-end education process*.

In this review article, we review **how machine learning and artificial intelligence can be utilized in different phases of the end-to-end education process—from planning and scheduling to knowledge delivery and assessment. To systematically identify the different areas of active research with respect to engagement of AI in education, we first introduce a broad categorization of research articles in literature into those that address tasks prior to knowledge delivery and those that are relevant during** the process of knowledge delivery—i.e., *proactive vs. reactive engagement* with education. Proactive involvement of AI in education comes from its use in student admission logistics, curriculum design, scheduling and teaching content generation. Reactive involvement of AI is considerably broader in scope—AI-based methods can be used for designing intelligent tutoring systems, assessing performance and predicting student outcomes. In the schematic in Figure 1, we present an overview of our categorization approach. We have selected a sample set of research articles under each category and identified the key problem statements addressed using AI methods in the past 20 years. We believe that our categorization approach exposes to researchers the wide scope of using AI for the educational process. At the same time, it allows readers to identify the timeline of when certain AI-driven tool might be applicable and what are the key challenges and concerns with using these tools at that time. The article further summarizes for expert researchers how the use of datasets and algorithms have evolved over the years and the scope for future research in this domain.

Through this review article, we aim to address the following questions:

- What were the widely studied applications of artificial intelligence in the end-to-end education process in the past two decades? How did the 2020 outbreak of the COVID-19 pandemic influence the landscape of research in this domain? Over the past two decades in retrospective view, has the usage of AI for education widened or bridged the gap between population groups with respect to access to quality education?
- How has the choice of datasets and algorithms in AI-driven tools and platforms evolved over this period—particularly in addressing the active research questions in the end-to-end education process?

The organization of this review article from here on is as follows. In Section 2, we define the scope of this review, outline the paper selection strategy and present the summary statistics. In Section 3, we contextualize our contribution in the light of technical review articles published in the domain of AIEd in the past 5 years. In Section 4, we present our categorization approach and review the scientific and technical contributions in each category. Finally, in Section 5, we discuss the major trends observed in research in the AIEd sector over the past two decades, discuss how the COVID-19 pandemic is reshaping the AIEd landscape and point out existing limitations in the global adoption of AI-driven tools for education. Additionally in Table 1, we provide a glossary of technical terms and their abbreviations that have been used throughout the paper.

2. Scope definition

The term artificial intelligence (AI) was coined in 1956 by John McCarthy (Haenlein and Kaplan, 2019). Since the first generally acknowledged work of McCulloch and Pitts in conceptualizing artificial neurons, AI has gone through several dormant periods and shifts in research focus. From algorithms that through exposure to somewhat noisy observational data learns to perform some pre-defined tasks, i.e., *machine learning (ML)* to more sophisticated approaches that learns the mapping of high-dimensional observations to representations in a lower dimensional space, i.e., *deep learning (DL)*—there is a plethora of computational techniques available currently. More recently, researchers and social scientists are increasingly using AI-based techniques to address social issues and to build toward a sustainable future (Shi et al., 2020). In this article, we focus on how one such social development aspect, i.e., education might benefit from usage of artificial intelligence, machine learning, and deep learning methods.

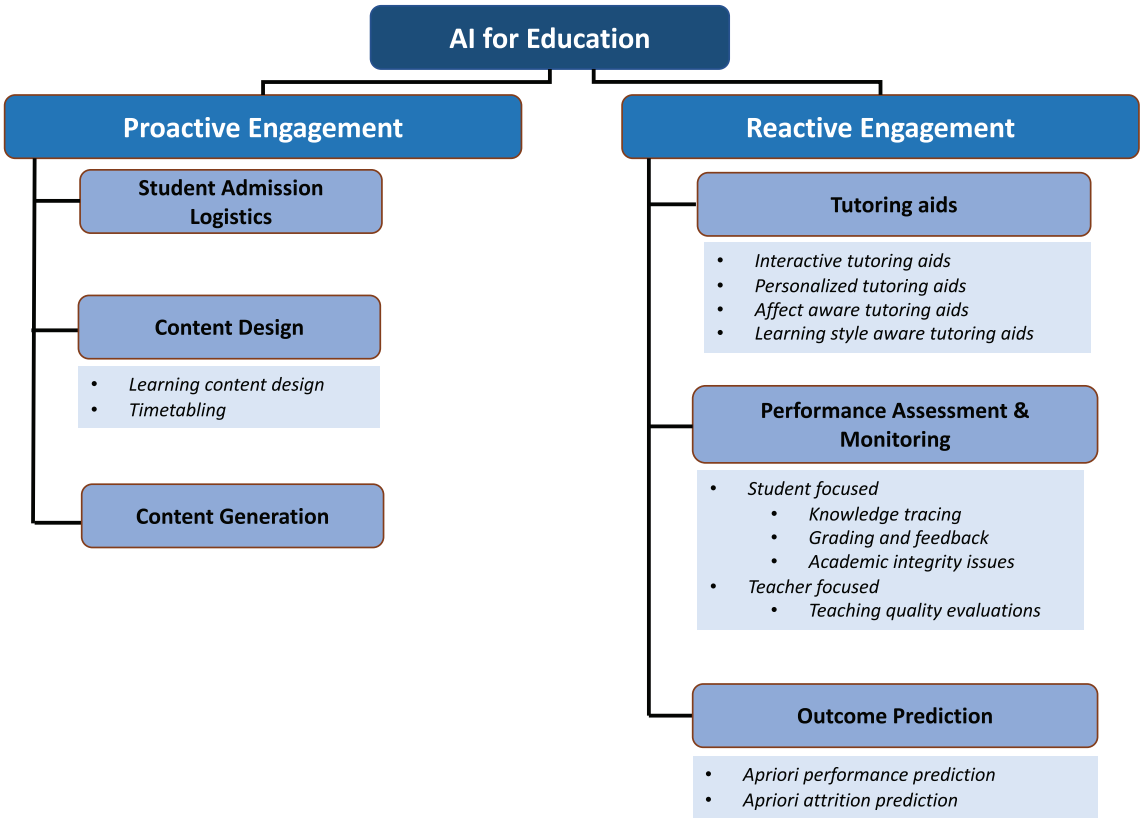


FIGURE 1
Overview of the categorization introduced in this review article.

2.1. Paper search strategy

For the purpose of analyzing recent trends in this field (i.e., AIEd), we have sampled research articles published in peer-reviewed conferences and journals over the past 20 years, i.e. between 2003 and 2022, by leveraging the Google Scholar search engine. We identified our selected corpus of 195 research articles through a multi-step process. First, we identified a set of systematic review, survey papers and perspective papers published in the domain of artificial intelligence for education (AIEd) between the years of 2018 and 2022. To identify this list of review papers we used the keywords "artificial intelligence for education", "artificial intelligence for education review articles" and similar combinations in Google Scholar. We critically reviewed these papers and identified the research domains under AIEd that have received much attention in the past 20 years (i.e., 2002–2022) and that are closely tied to the end-to-end education process. Once, these research domains were identified, we further did a deep dive search using relevant keywords for each research area (for example, for the category tutoring aids, we used several keywords including intelligent tutoring systems, intelligent tutoring aids, computer-aided learning systems, affect-aware learning systems) to identify an initial set of technical papers in the sub-domain. We streamlined this initial set through the lens of significance of the problem statement, data used, algorithm proposed by thorough review of each paper by both authors and retained the final set of 195 research articles.

2.2. Inclusion and exclusion criteria

Since the coinage of the term artificial intelligence, there is considerable debate in the scientific community about what is the scope of artificial intelligence. It is specifically challenging to delineate the boundaries as it is indeed a field that is subject to rapid technological change. Deep-dive analysis of this debate is beyond the scope of this paper. Instead, we have clearly stated in this section our inclusion/exclusion criteria with respect to selecting articles that surfaced in our search of involvement of AI for education. For this review article, we include research articles that use methods such as optimal search strategies (e.g., breadth-first search, depth-first search), density estimation, machine learning, Bayesian machine learning, deep learning and reinforcement learning. We do not include original research that proposes use of concepts and methods rooted in operations research, evolutionary algorithms, adaptive control theory, and robotics in our corpus of selected articles. In this review, we **only** consider peer-reviewed articles that were published in English. We do not include patented technologies and copyrighted EdTech software systems in our scope unless

peer-reviewed articles outlining the same contributions have been published by the authors.

2.3. Summary statistics

With the scope of our review defined above, here we provide the summary statistics of the 195 technical articles we covered in this review. In Figure 2, we show the distribution of the included scientific and technical articles over the past two decades. We also introspected the technical contributions in each category of our categorization approach with respect to the target audiences they catered to (see Figure 3). We primarily identify target audience groups for educational technologies as such—pre-school students, elementary school students, middle and high school students, university students, standardized test examinees, students in e-learning platforms, students of MOOCs, and students in professional/vocational education. Articles where the audience group has not been clearly mentioned were marked as belonging to "Unknown" target audience category.

In Section 4, we introduce our categorization and perform a deep-dive to explore the breadth of technical contributions in each category. If applicable, we have further identified specific research problems currently receiving much attention as sub-categories within a category. In Table 2, we demonstrate the distribution of significant research problems within a category.

We defer the analysis of the identified trends from these summary plots to the Section 5 of this paper.

TABLE 1 Glossary of technical terms and their abbreviations frequently used in the paper.

Artificial Intelligence (AI): Simulation of human intelligence processes by machines.
Machine Learning (ML): Technologies or algorithms enabling computer systems to identify patterns from data, make decisions and improve their performance through experience.
Bayesian Machine Learning: A paradigm for constructing statistical models based on Bayes Theorem.
Deep Learning (DL): A class of machine learning algorithms that uses artificial neural networks consisting of multiple processing layers to map raw data into progressively higher-dimensional features.
Supervised Learning: A type of machine learning problem where algorithms are trained using labeled data points for the purpose of predicting labels for unseen examples.
Unsupervised Learning: A type of machine learning problem that learns patterns from unlabeled data.
Reinforcement Learning (RL): A type of machine learning problem where an agent learns an optimal set of actions in an environment through trial and feedback in order to maximize a reward.
Natural Language Processing (NLP): A branch of artificial intelligence and machine learning that enables computer systems to process and analyze natural language data in written or spoken format.
Convolutional Neural Networks (CNNs): A type of artificial neural network consisting of convolutional layers, most commonly used for processing visual imagery.
Generative AI: A type of artificial intelligence technology based on generative models that can produce text, images, audio and other kinds of content in response to prompts.

3. Related works

Artificial intelligence as a research area in technology has evolved gradually since 1950s. Similarly, the field of using computer based technology to support education has been actively developing since the 1980s. It is only however in the past few decades that there has been significant emphasis in adopting digital technologies including AI driven technologies in practice (Alam, 2021). Particularly, the introduction of open source generative AI algorithms, has spear-headed critical analyses of how AI can and should be used in the education sector (Baidoo-Anu and Owusu Ansah, 2023; Lund and Wang, 2023). In this backdrop of emerging developments, the number of review articles surveying the technical progress in the AIEd discipline has also increased in the last decade (see Figure 4). To generate Figure 4, we used Google Scholar as the search engine with the keywords artificial intelligence for education, artificial intelligence for education review articles and similar combinations using domain abbreviations. In this section, we discuss the premise of the review articles published in the *last 5*

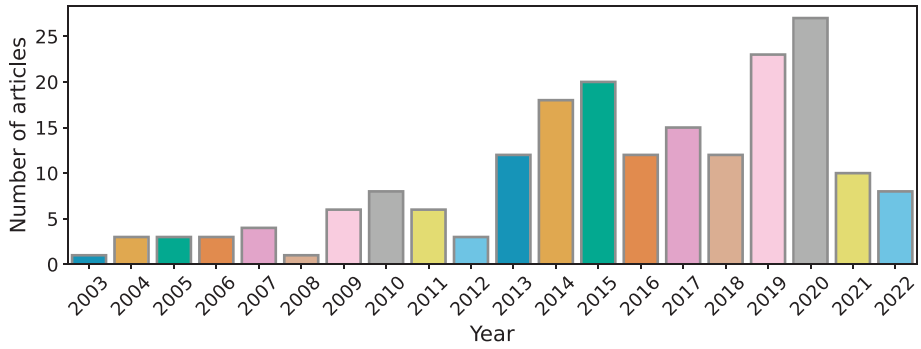


FIGURE 2
Distribution of the reviewed technical articles across the past two decades.

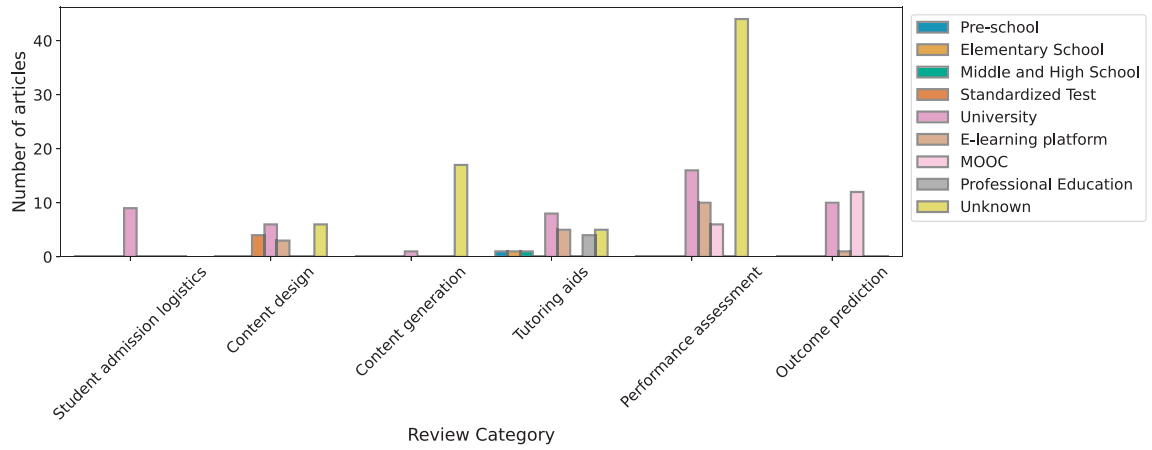


FIGURE 3
Distribution of reviewed technical articles across categories and target audience categories.

TABLE 2 Distribution of reviewed technical articles across sub-categories under each category.

Proactive vs. reactive engagement of AI	Review category	Review subcategory	Count
Proactive engagement	Student admission logistics	N/A	9
Proactive engagement	Content design	Learning content design	15
Proactive engagement	Content design	Timetabling	4
Proactive engagement	Content generation	N/A	22
Reactive engagement	Tutoring aids	Interactive tutoring aids	13
Reactive engagement	Tutoring aids	Personalized tutoring aids	8
Reactive engagement	Tutoring aids	Learning style based tutoring aids	7
Reactive engagement	Tutoring aids	Affect aware tutoring aids	5
Reactive engagement	Performance assessment	Student-focused	76
Reactive engagement	Performance assessment	Teacher-focused	9
Reactive engagement	Outcome prediction	Performance prediction	13
Reactive engagement	Outcome prediction	Drop-out prediction	14

years and situate this article with respect to previously published technical reviews.

Among the review articles identified based on the keyword search on Google Scholar and published between 2018 and 2022, one can identify two thematic categories—(i) *Technical reviews with categorization*: review articles that group research contributions based on some distinguishing factors, such as problem statement and solution methodology (Chassignol et al., 2018; Zawacki-Richter et al., 2019; Ahmad et al., 2020, 2022; Chen L. et al., 2020; Yufeia et al., 2020; Huang J. et al., 2021; Lameras and Arnab, 2021; Ouyang and Jiao, 2021; Zhai et al., 2021; Chen et al., 2022; Holmes and Tuomi, 2022; Namatherdhala et al., 2022; Wang and Cheng, 2022). (ii) *Perspectives on challenges, trends, and roadmap*: review articles that highlight the current state of research in a domain and offer critical analysis of the challenges and the future road map for the domain (Fahimirad and Kotamjani, 2018; Humble and Mozelius, 2019; Malik et al., 2019; Pedro et al., 2019; Bryant et al., 2020; Hwang et al., 2020; Alam, 2021; Schiff, 2021). Closely linked with (i) are review articles that dive deep into the developments within a particular sub-category associated with AIEd, such as AIEd in the context of early childhood education (Su and Yang, 2022) and online higher education (Ouyang F. et al., 2022). We have designed this review article to belong to category (i). We distinguish between the different research problems in the context of AIEd through the lens of their timeline for engagement in the end-to-end education process and then perform a deeper review of ongoing research efforts in each category. To the best of our knowledge, such distinction between proactive and reactive involvement of AI in education along with an granular review of significant research questions in each category is presented for the first time through this paper (see schematic in Figure 1).

In Table 3, we have outlined the context of recently published technical reviews with categorization.

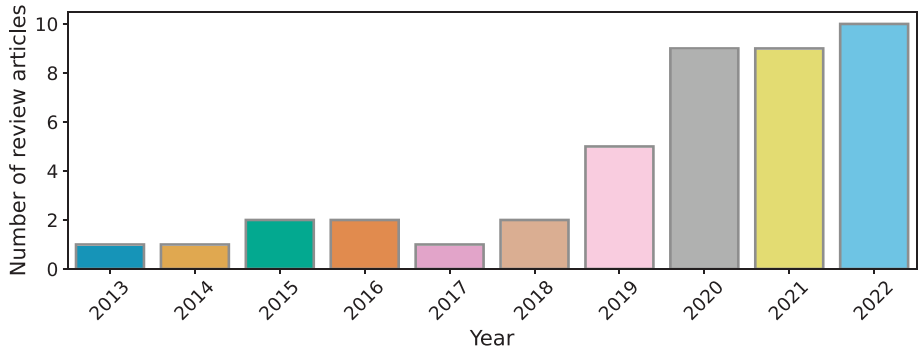


FIGURE 4
Number of review articles published in AIEd over the past decade.

4. Engaging artificial intelligence driven methods in stages of education

4.1. Proactive vs. reactive engagement of AI—An introduction

In the introductory section of this article, we have outlined how the process of education is a multi-step process and how it involves different stakeholders along the timeline. To this end, we can clearly identify that there are two distinct phases of engaging AI in the end-to-end education process. First, **proactive engagement** of AI—efforts in this phase are to design, curate and to ensure optimal use of resources, and second, **reactive engagement** of AI—efforts in this phase are to ensure that students acquire the necessary information and skills from the sessions they attend and provide feedback as needed.

In this review article, we distinguish between the scientific and technical contributions in the field of AIEd through the lens of these two distinct phases. This categorization is significant for the following reasons:

- First, through this hierarchical categorization approach, one can gauge the range of problems in the context of education that can be addressed using artificial intelligence. AI research related to personalized tutoring aids and systems has indeed had a head-start and is a mature area of research currently. However, the scope of using AI in the end-to-end education process is broad and rapidly evolving.
- Second, this categorization approach provides a retrospective overview of milestones achieved in AIEd through continuous improvement and enrichment of the data and algorithm leveraged in building AI models.
- Third, as this review touches upon both classroom and administrative aspect of education, readers can formulate a perspective for the myriad of infrastructural and ethical challenges that exist with respect to widespread adoption of AI-driven methods in education.

Within these broad categorizations, we further break down and analyze the research problems that have been addressed using AI. For instance, in the proactive engagement phase, AI-based algorithms can be leveraged to determine student admission logistics, design curricula and schedules, and create course content. On the other hand, in the reactive engagement phase, AI-based methods can be used for designing intelligent tutoring systems (ITS), performance assessment, and prediction of student outcomes (see Figure 1). Another important distinction between the two phases lies in the nature of the available data to develop models. While the former primarily makes use of historical data points or pre-existing estimates of available resources and expectations about learning outcomes, the latter has at its disposal a growing pool of data points from the currently ongoing learning process, and can therefore be more adaptive and initiate faster pedagogical interventions to changing scopes and requirements.

4.2. Proactive engagement of AI for education

4.2.1. Student admission logistics

In the past, although a number of studies used statistical or machine learning-based approaches to analyze or model student admissions decisions, they had little role in the actual admissions process (Bruggink and Gambhir, 1996; Moore, 1998). However in the face of growing numbers of applicants, educational institutes are increasingly turning to AI-driven approaches to efficiently review applications and make admission decisions. For example, the Department of Computer Science at University of Texas Austin (UTCS) introduced an explainable AI system called GRADE (Graduate Admissions Evaluator) that uses logistic regression on past admission records to estimate the probability of a new applicant being admitted in their graduate program (Waters and Miikkulainen, 2014). While GRADE did not make the final admission decision, it reduced the number of full application reviews as well as review time per application by experts. Zhao et al. (2020) used features extracted from application materials of students as well as how they performed in the program of study to predict an incoming applicant's potential performance and identify students best suited for the program. An important metric for educational institutes with regard to student admissions

TABLE 3 Contextualization with respect to technical reviews published in the past 5 years (2018–2022).

Paper title	Summary
Artificial Intelligence trends in education: a narrative overview (Chassignol et al., 2018)	Categorizes AI in education into four categories—customized educational content, assessment and evaluation, adaptive systems and personalization, intelligent tutoring systems.
Systematic review of research on artificial intelligence applications in higher education—where are the educators (Zawacki-Richter et al., 2019)	Categorizes AI in education into four categories—profiling and prediction, assessment and evaluation, adaptive systems and personalization, intelligent tutoring systems.
Artificial Intelligence in Education: A Review (Chen L. et al., 2020)	Identifies and reviews four key ways in which AI has been adopted for education—automation of administrative processes and tasks, curriculum and content development, instruction, modeling students' learning process.
Review of the application of artificial intelligence in education (Yufeia et al., 2020)	Identifies and reviews aspects in which AI technology has been used in education—automatic grading system, interval reminder, teacher's feedback, virtual teachers, personalized learning, adaptive learning, augmented reality/virtual reality, accurate reading, intelligent campus and distance learning.
Artificial Intelligence in Education: A panoramic review (Ahmad et al., 2020)	Reviews the various applications of AI such as student grading and evaluations, students retention and drop out prediction, sentiment analysis, intelligent tutoring, classroom monitoring and recommendation systems.
A Review of Artificial Intelligence (AI) in Education from 2010 to 2020 (Zhai et al., 2021)	Reviews articles that use AI for social sciences such as in education and classifies the research questions into development layer (classification, matching, recommendation, and deep learning), application layer (feedback, reasoning, and adaptive learning), and integration layer (affection computing, role-playing, immersive learning, and gamification).
Artificial intelligence in education: the three paradigms (Ouyang and Jiao, 2021)	Identifies the paradigm shifts of AIEd and categorizes into AI-directed (learner-as-recipient), AI-supported (learner-as-collaborator), and AI-empowered (learner-as-leader).
Power to the teachers: an exploratory review on artificial intelligence in education (Lameras and Arnab, 2021)	Discusses research contribution along the five aspects of teaching and learning introduced by Dong and Chen (2020): 1. AIEd for preparing and transmitting learning content 2. AIEd for helping students to apply knowledge 3. AIEd for engaging students in learning tasks 4. AIEd for helping students to improvement through assessments and feedback 5. AIEd for helping students to become self-regulated learners.
A review on artificial intelligence in education (Huang J. et al., 2021)	Outlines the application of AI in education—adaptive learning, teaching evaluation, virtual classroom, smart campus, intelligent tutoring robots, and then analyzes its impact on teaching and learning.
Toward a tripartite research agenda: a scoping review of artificial intelligence in education research (Wang and Cheng, 2022)	Provides a scoping review of research studies on AIEd published between 2001 and 2021 and identifies and discusses three distinct agendas—Learning from AI, Learning about AI, and Learning with AI.
Two Decades of Artificial Intelligence in Education: contributors, Collaborations, Research Topics, Challenges, and Future Directions (Chen et al., 2022)	The authors identify the main research topics in AIEd in the past two decades to be—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.
Academic and Administrative Role of Artificial Intelligence in Education (Ahmad et al., 2022)	This review article aims to explore the academic and administrative applications of AI with an in-depth discussion on artificial intelligence applications in 1. Grading/Assessment 2. Admission 3. Virtual Reality (VR) for education 4. Learning Analytics.
A Comprehensive Overview of Artificial Intelligence Trends in Education (Namatherdhala et al., 2022)	The authors categorize application of AI for education into three distinct groups—Education administration, Instruction Design and Learning outcomes and briefly reviews each of them.
State of the art and practice in AI in education (Holmes and Tuomi, 2022)	The authors provide a review of existing AI systems in education and their pedagogic and educational assumptions. They also introduce a categorization for AIEd systems and discusses different ways of using AI in education and learning and different interpretations of what AI and education is or could be and existing roadblocks.

is yield rate, the rate at which accepted students decide to enroll at a given school. Machine learning has been used to predict enrollment decisions of students, which would help the institute make strategic admission decisions in order to improve their yield rate and optimize resource allocation (Jamison, 2017). Additionally, whether students enroll in suitable majors based on their specific backgrounds and prior academic performance is also indicative of future success. Machine learning has also been used to classify students into suitable majors in an attempt to set them up for academic success (Assiri et al., 2022).

Another research direction in this domain approaches the admissions problem from the perspective of students by predicting the probability that an applicant will get admission at a particular university in order to help applicants better target universities based on their profiles as well as university rankings (AlGhamdi et al., 2020; Goni et al., 2020; Mridha et al., 2022). Notably, more than one such work finds prior GPA (Grade Point Average) of students to be the most significant factor in admissions decisions (Young and Caballero, 2019; El Guabassi et al., 2021).

Given the high stakes involved and the significant consequences that admissions decisions have on the future of students, there has been considerable discourse on the ethical considerations of using AI in such applications, including its fairness, transparency, and privacy aspects (Agarwal, 2020; Finocchiaro et al., 2021). Aside from the obvious potential risks of worthy applicants getting rejected or unworthy applicants getting in, such systems

can perpetuate existing biases in the training data from human decision-making in the past (Bogina et al., 2022). For example, such systems might show unintentional bias toward certain demographics, gender, race, or income groups. Bogina et al. (2022) advocated for explainable models for making admission decisions, as well as proper system testing and balancing before reaching the end user. Emelianov et al. (2020) showed that demographic parity mechanisms like group-specific admission thresholds increase the utility of the selection process in such systems in addition to improving its fairness. Despite concerns regarding fairness and ethics, interestingly, university students in a recent survey rated algorithmic decision-making (ADM) higher than human decision-making (HDM) in admission decisions in both procedural and distributive fairness aspects (Marcinkowski et al., 2020).

4.2.2. Content design

In the context of education, we can define content as—(i) learning content for a course, curriculum, or test; and (ii) schedules/timetables of classes. We discuss AI/ML approaches for designing/structuring both of the above in this section.

(i) **Learning content design**: Prior to the start of the learning process, educators, and administrators are responsible for identifying an appropriate set of courses for a curriculum, an appropriate set of contents for a course, or an appropriate set of questions for a standardized test. In course and curriculum design, there is a large body of work using traditional systematic and relational approaches (Kessels, 1999), however the last decade saw several works using AI-informed curriculum design approaches. For example, Ball et al. (2019) uses classical ML algorithms to identify factors prior to declaration of majors in universities that adversely affect graduation rates, and advocates curriculum changes to alleviate these factors. Rawatlal (2017) uses tree-based approaches on historical records to prioritize the prerequisite structure of a curriculum in order to determine student progression routes that are effective. Somasundaram et al. (2020) proposes an Outcome Based Education (OBE) where expected outcomes from a degree program such as job roles/skills are identified first, and subsequently courses required to reach these outcomes are proposed by modeling the curriculum using ANNs. Doroudi (2019) suggests a semi-automated curriculum design approach by automatically curating low-cost, learner-generated content for future learners, but argues that more work is needed to explore data-driven approaches in curating pedagogically useful peer content.

For designing standardized tests such as TOEFL, SAT, or GRE, an essential criteria is to select questions having a consistent difficulty level across test papers for fair evaluation. This is also useful in classroom settings if teachers want to avoid plagiarism issues by setting multiple sets of test papers, or in designing a sequence of assignments or exams with increasing order of difficulty. This can be done through Question Difficulty Prediction (QDP) or Question Difficulty Estimation (QDE), an estimate of the skill level needed to answer a question correctly. QDP was historically estimated by pretesting on students or from expert ratings, which are expensive, time-consuming, subjective, and often vulnerable to leakage or exposure (Benedetto et al., 2022). Rule-based algorithms relying on difficulty features extracted by experts were also proposed in Grivokostopoulou et al. (2014) and Perikos et al. (2016) for automatic difficulty estimation. As data-driven solutions became more popular, a common approach used linguistic features (Mothe and Tanguy, 2005; Stiller et al., 2016), readability scores, (Benedetto et al., 2020a; Yaneva et al., 2020), and/or word frequency features (Benedetto et al., 2020a,b; Yaneva et al., 2020) with ML algorithms such as linear regression, SVMs, tree-based approaches, and neural networks for downstream classification or regression, depending on the problem setup. With automatic testing systems and ready availability of large quantities of historical test logs, deep learning has been increasingly used for feature extraction (word embeddings, question representations, etc.) and/or difficulty estimation (Fang et al., 2019; Lin et al., 2019; Xue et al., 2020). Attention strategies have been used to model the difficulty contribution of each sentence in reading problems (Huang et al., 2017) or to model recall (how hard it is to recall the knowledge assessed by the question) and confusion (how hard it is to separate the correct answer from distractors) in Qiu et al. (2019). Domain adaptation techniques have also been proposed to alleviate the need of difficulty-labeled question data for each new course by aligning it with the difficulty distribution of a resource-rich course (Huang Y. et al., 2021). AlKhuzaey et al. (2021) points out that a majority of data-driven QDP approaches belong to language learning and medicine, possibly spurred on by the existence of a large number of international and national-level standardized language proficiency tests and medical licensing exams.

(ii) **Timetabling**: Educational Timetabling Problem (ETP) deals with the assignment of classes or exams to a limited number of time-slots such that certain constraints (e.g., availability of teachers, students, classrooms, and equipments) are satisfied. This can be divided into three types—course timetabling, school timetabling, and exam timetabling (Zhu et al., 2021). Timetabling not only ensures proper resource allocation, its design considerations (e.g., number of courses per semester, number of lectures per day, number of free time-slots per day) have noticeable impact on student attendance behavior and academic performance (Larabi-Marie-Sainte et al., 2021). Popular approaches in this domain such as mathematical optimization, meta-heuristic, hyper-heuristic, hybrid, and fuzzy logic approaches. Zhu et al. (2021) and Tan et al. (2021) mostly is beyond the scope of our paper (see Section 2.2). Having said that, it must be noted that machine learning has often been used in conjunction with such mathematical techniques to obtain better performing algorithms. For example, Kenekayoro (2019) used supervised learning to find approximations for evaluating solutions to optimization problems—a critical step in heuristic approaches. Reinforcement learning has been used to select low-level heuristics in hyper-heuristic approaches (Obit et al., 2011; Özcan et al., 2012) or to obtain a suitable search neighborhood in mathematical optimization problems (Goh et al., 2019).

4.2.3. Content generation

The difference between content design and content generation is that of curation versus creation. While the former focuses on selecting and structuring the contents for a course/curriculum in a

way most appropriate for achieving the desired learning outcomes, the latter deals with generating the course material itself. AI has been widely adopted to generate and improve learning content prior to the start of the learning process, as discussed in this section.

Automatically generating questions from narrative or informational text, or automatically generating problems for analytical concepts are becoming increasingly important in the context of education. Automatic question generation (AQG) from teaching material can be used to improve learning and comprehension of students, assess information retention from the material and aid teachers in adding Supplementary material from external sources without the time-intensive process of authoring assessments from them. They can also be used as a component in intelligent tutoring systems to drive engagement and assess learning. AQG essentially consists of two aspects: content selection or *what to ask*, and question construction or *how to ask it* (Pan et al., 2019), traditionally considered as separate problems. Content selection for questions was typically done using different statistical features (sentence length, word/sentence position, word frequency, noun/pronoun count, presence of superlatives, etc.) (Agarwal and Mannem, 2011) or NLP techniques such as syntactic or semantic parsing (Heilman, 2011; Lindberg et al., 2013), named entity recognition (Kalady et al., 2010) and topic modeling (Majumder and Saha, 2015). Machine learning has also been used in such contexts, e.g., to classify whether a certain sentence is suitable to be used as a stem in cloze questions (passage with a portion occluded which needs to be replaced by the participant) (Correia et al., 2012). The actual question construction, on the other hand, traditionally adopted rule-based methods like transformation-based approaches (Varga and Ha, 2010) or template-based approaches (Mostow and Chen, 2009). The former rephrased the selected content using the correct question key-word after deleting the target concept, while the latter used pre-defined templates that can each capture a class of questions. Heilman and Smith (2010) used an overgenerate-and-rank approach to overgenerate questions followed by the use of supervised learning for ranking them, but still relied on handcrafted generating rules. Following the success of neural language models and concurrent with the release of large-scale machine reading comprehension datasets (Nguyen et al., 2016; Rajpurkar et al., 2016), question generation was later framed as a sequence-to-sequence learning problem that directly maps a sentence (or the entire passage containing the sentence) to a question (Du et al., 2017; Zhao et al., 2018; Kim et al., 2019), and can thus be trained in an end-to-end manner (Pan et al., 2019). Reinforcement learning based approaches that exploit the rich structural information in the text have also been explored in this context (Chen Y. et al., 2020). While text is the most common type of input in AQG, such systems have also been developed for structured databases (Jouault and Seta, 2013; Indurthi et al., 2017), images (Mostafazadeh et al., 2016), and videos (Huang et al., 2014), and are typically evaluated by experts on the quality of generated questions in terms of relevance, grammatical, and semantic correctness, usefulness, clarity etc.

Automatically generating problems that are similar to a given problem in terms of difficulty level, can greatly benefit teachers in setting individualized practice problems to avoid plagiarism and still ensure fair evaluation (Ahmed et al., 2013). It also enables the students to be exposed to as many (and diverse) training exercises as needed in order to master the underlying concepts (Keller, 2021). In this context, mathematical word problems (MWPs)—an established way of inculcating math modeling skills in K-12 education—have witnessed significant research interest. Preliminary work in automatic MWP generation take a template-based approach, where an existing problem is generalized into a template, and a solution space fitting this template is explored to generate new problems (Deane and Sheehan, 2003; Polozov et al., 2015; Koncel-Kedziorski et al., 2016). Following the same shift as in AQG, Zhou and Huang (2019) proposed an approach using Recurrent Neural Networks (RNNs) that encodes math expressions and topic words to automatically generate such problems. Subsequent research along this direction has focused on improving topic relevance, expression relevance, language coherence, as well as completeness and validity of the generated problems using a spectrum of approaches (Liu et al., 2021; Wang et al., 2021; Wu et al., 2022).

On the other end of the content generation spectrum lie systems that can generate solutions based on the content and related questions, which include Automatic Question Answering (AQA) systems, Machine Reading Comprehension (MRC) systems and automatic quantitative reasoning problem solvers (Zhang D. et al., 2019). These have achieved impressive breakthroughs with the research into large language models and are widely regarded in the larger narrative as a stepping-stone toward Artificial General Intelligence (AGI), since they require sophisticated natural language understanding and logical inferencing capabilities. However, their applicability and usefulness in educational settings remains to be seen.

4.3. Reactive engagement of AI for education

4.3.1. Tutoring aids

Technology has been used to aid learners to achieve their learning goals for a long time. More focused effort on developing computer-based tutoring systems in particular started following the findings of Bloom (Bloom, 1984)—students who received tutoring in addition to group classes fared two standard deviations better than those who only participated in group classes. Given its early start, research on Intelligent Tutoring Systems (ITS) is relatively more mature than other research areas under the umbrella of AIEd research. Fundamentally, the difference between designs of ITS comes from the difference in the *underlying assumption of what augments the knowledge acquisition process for a student*. In the review paper on ITS (Alkhatlan and Kalita, 2018), a comprehensive timeline and overview of research in this domain is provided. Instead of repeating findings from previous reviews under this category, we distinguish between ITS designs through the lens of the underlying hypotheses. We primarily identified four hypotheses that are currently receiving much attention from the research community—emphasis on tutor-tutee interaction, emphasis of personalization, inclusion of affect and emotion, and consideration of specific learning styles. It must be noted that tutoring itself is

an interactive process, therefore most designs in this category have a basic interactive setup. However, contributions in categories (ii) through (iv), have other concept as the focal point of their tutoring aid design.

(i) **Interactive tutoring aids**: Previous research in education (Jackson and McNamara, 2013) has pointed out that *when a student is actively interacting with the educator or the course contents, the student stays engaged in the learning process for a longer duration*. Learning systems that leverage this hypothesis can be categorized as interactive tutoring aids. These frameworks allow the student to communicate (verbally or through actions) with the teacher or the teaching entity (robots or software) and get feedback or instructions as needed.

Early designs of interactive tutoring aids for teaching and support comprised of rule-based systems mirroring interactions between expert teacher and student (Arroyo et al., 2004; Olney et al., 2012) or between peer companions (Movellan et al., 2009). These template rules provided output based on the inputs from the student. Over the course of time, interactive tutoring systems gradually shifted to inferring the student's state in real time from the student's interactions with the tutoring system and providing fine-tuned feedback/instructions based on the inference. For instance, Gordon and Breazeal (2015) used a Bayesian active learning algorithm to assess student's word reading skills while the student was being taught by a robot. Presently, a significant number of frameworks belonging to this category uses chatbots as a proxy for a teacher or a teaching assistant (Ashfaque et al., 2020). These recent designs can use a wide variety of data such as text and speech, and rely on a combination of sophisticated and resource-intensive deep-learning algorithms to infer and further customize interactions with the student. For example, Pereira (2016) presents "@dawebot" that uses NLP techniques to train students using multiple choice question quizzes. Afzal et al. (2020) presents a conversational medical school tutor that uses NLP and natural language understanding (NLU) to understand user's intent and present concepts associated with a clinical case.

Hint construction and partial solution generation is yet another method to keep students engaged interactively. For instance, Green et al. (2011) used Dynamic Bayes Nets to construct a curriculum of hints and associated problems. Wang and Su (2015) in their architecture iGeoTutor assisted students in mastering geometry theorems by implementing search strategies (e.g., DFS) from partially complete proofs. Pande et al. (2021) aims to improve individual and self-regulated learning in group assignments through a conversational system built using NLU and dialogue management systems that prompts the students to reflect on lessons learnt while directing them to partial solutions.

One of the requirements of certain professional and vocational training such as biology, medicine, military etc. is practical experience. With the support of booming infrastructure, many such training programs are now adopting AI-driven augmented reality (AR)/virtual reality (VR) lesson plans. Interconnected modules driven by computer vision, NLU, NLP, text-to-speech (TTS), information retrieval algorithms facilitate lessons and/or assessments in biology (Ahn et al., 2018), surgery and medicine (Mirchi et al., 2020), pathological laboratory analysis (Taoum et al., 2016), and military leadership training (Gordon et al., 2004).

(ii) **Personalized tutoring aids**: As every student is unique, *personalizing instruction and teaching content can positively impact the learning outcome of the student* (Walkington, 2013)—tutoring systems that incorporate this can be categorized as personalized learning systems or personalized tutoring aids. Notably, personalization during instruction can occur through course content sequencing and display of prompts and additional resources among others.

The sequence in which a student reviews course topics plays an important role in their mastery of a concept. One of the criticisms of early computer based learning tools was the "one approach fits all" method of execution. To improve upon this limitation, personalized instructional sequencing approaches were adopted. In some early developments, Idris et al. (2009) developed a course sequencing method that mirrored the role of an instructor using soft computing techniques such as self organized maps and feed-forward neural networks. Lin et al. (2013) propose the use of decision trees trained on student background information to propose personalized learning paths for creativity learning. Reinforcement learning (RL) naturally lends itself to this task. Here an optimal policy (sequence of instructional activities) is inferred depending on the cognitive state of a student (estimated through knowledge tracing) in order to maximize a learning-related reward function. As knowledge delivery platforms are increasingly becoming virtual and thereby generating more data, deep reinforcement learning has been widely applied to the problem of instructional sequencing (Reddy et al., 2017; Upadhyay et al., 2018; Pu et al., 2020; Islam et al., 2021). Doroudi (2019) presents a systematic review of RL-induced instructional policies that were evaluated on students, and concludes that over half outperform all baselines they were tested against.

In order to display a set of relevant resources personalized with respect to a student state, algorithmic search is carried out in a knowledge repository. For instance, Kim and Shaw (2009) uses information retrieval and NLP techniques to present two frameworks: PedaBot that allows students to connect past discussions to the current discussion thread and MentorMatch that facilitates student collaboration customized based on student's current needs. Both PedaBot and MentorMatch systems use text data coming from a live discussion board in addition to textbook glossaries. In order to reduce information overload and allow learners to easily navigate e-learning platforms, Deep Learning-Based Course Recommender System (DECOR) has been proposed recently (Li and Kim, 2021)—this architecture comprises of neural network based recommendation systems trained using student behavior and course related data.

(iii) **Affect aware tutoring aids**: Scientific research proposes *incorporating affect and behavioral state of the learner into the design of the tutoring system as it enhances the effectiveness of the teaching process* (Woolf et al., 2009; San Pedro et al., 2013). Arroyo et al. (2014) suggests that cognition, meta-cognition and affect should indeed be modeled using real time data and used to design intervention strategies. Affect and behavioral state of a student can generally be inferred from sensor data that tracks minute physical movements of the student (eyegaze, facial expression, posture etc.). While initial approaches in this direction required sensor data, a major constraint for availing and using such data pertains to

ethical and legal reasons. "Sensor-free" approaches have thereby been proposed that use data such as student self-evaluations and/or interaction logs of the student with the tutoring system. Arroyo et al. (2010) and Woolf et al. (2010) use interaction data to build affect detector models—the raw data in these cases are first distilled into meaningful features and then fed into simple classifier models that detect individual affective states. DeFalco et al. (2018) compares the usage of sensor and interaction data in delivering motivational prompts in the course of military training. In Botelho et al. (2017), uses RNNs to enhance the performance of sensor-free affect detection models. In their review of affect and emotion aware tutoring aids, Harley et al. (2017) explore in depth the different use cases for affect aware intelligent tutoring aids such as enriching user experience, better curating learning material and assessments, delivering prompts for appraisal, navigational instructions etc., and the progress of research in each direction.

(iv) **Learning style aware tutoring aids**: Yet another perspective in the domain of ITS *pertains to customizing course content according to learning styles of students for better end outcomes*. Kolb (1976), Pask (1976), Honey and Mumford (1986), and Felder (1988) among others proposed different approaches to categorize learning styles of students. Traditionally, an individual's learning style was inferred via use of a self-administered questionnaire. However, more recently machine learning based methods are being used to categorize learning styles more efficiently from noisy subject data. Lo and Shu (2005), Villaverde et al. (2006), Alfaro et al. (2018), and Bajaj and Sharma (2018) use as input the completed questionnaire and/or other data sources such as interaction data and behavioral data of students, and feed the extracted features into feed-forward neural networks for classification. Unsupervised methods such as self-organizing map (SOM) trained using curated features have also been used for automatic learning style identification (Zatarain-Cabada et al., 2010). While for categorization per the Felder and Silverman learning style model, count of student visits to different sections of the e-learning platform are found to be more informative (Bernard et al., 2015; Bajaj and Sharma, 2018), for categorization per the Kolb learning model, student performance, and student preference features were found to be more relevant. Additionally, machine learning approaches have also been proposed for learning style based learning path design. In Mota (2008), learning styles are first identified through a questionnaire and represented on a polar map, thereafter neural networks are used to predict the best presentation layout of the learning objective for a student. It is worthwhile to point out, however, that in recent years instead of focusing on customizing course content with respect to certain pre-defined learning styles, more research efforts are focused on curating course material based on how an individual's overall preferences vary over time (Chen and Wang, 2021).

4.3.2. Performance assessment and monitoring

A critical component of the knowledge delivery phase involves assessing student performance by tracing their knowledge development and providing grades and/or constructive feedback on assignments and exams, while simultaneously ensuring academic integrity is upheld. Conversely, it is also important to evaluate the quality and effectiveness of teaching, which has a tangible impact on the learning outcomes of students. AI-driven performance assessment and monitoring tools have been widely developed for both learners and educators. Since a majority of evaluation material are in textual format, NLP-based models in particular have a major presence in this domain. We divide this section into student-focused and teacher-focused approaches, depending on the direct focus group of such applications.

(i) **Student-focused**:

Knowledge tracing. An effective way of monitoring the learning progress of students is through knowledge tracing, which models knowledge development in students in order to predict their ability to answer the next problem correctly given their current mastery level of knowledge concepts. This not only benefits the students by identifying areas they need to work on, but also the educators in designing targeted exercises, personalized learning recommendations and adaptive teaching strategies (Liu et al., 2019). An important step of such systems is cognitive modeling, which models the latent characteristics of students based on their current knowledge state. Traditional approaches for cognitive modeling include factor analysis methods which estimate student knowledge by learning a function (logistic in most cases) based on various factors related to the students, course materials, learning and forgetting behavior, etc. (Pavlik and Anderson, 2005; Cen et al., 2006; Pavlik et al., 2009). Another research direction explores Bayesian inference approaches that update student knowledge states using probabilistic graphical models like Hidden Markov Model (HMM) on past performance records (Corbett and Anderson, 1994), with substantial research being devoted to personalizing such model parameters based on student ability and exercise difficulty (Yudelson et al., 2013; Khajah et al., 2014). Recommender system techniques based on matrix factorization have also been proposed, which predict future scores given a student-exercise performance matrix with known scores (Thai-Nghe et al., 2010; Toscher and Jahrer, 2010). Abdelrahman et al. (2022) provides a comprehensive taxonomy of recent work in deep learning approaches for knowledge tracing. Deep knowledge tracing (DKT) was one of the first such models which used recurrent neural network architectures for modeling the latent knowledge state along with its temporal dynamics to predict future performance (Piech et al., 2015a). Extensions along this direction include incorporating external memory structures to enhance representational power of knowledge states (Zhang et al., 2017; Abdelrahman and Wang, 2019), incorporating attention mechanisms to learn relative importance of past questions in predicting current response (Pandey and Karypis, 2019; Ghosh et al., 2020), leveraging textual information from exercise materials to enhance prediction performance (Su et al., 2018; Liu et al., 2019) and incorporating forgetting behavior by considering factors related to timing and frequency of past practice opportunities (Nagatani et al., 2019; Shen et al., 2021). Graph neural network based architectures were recently proposed in order to better capture dependencies between knowledge concepts or between questions and their underlying knowledge concepts (Nakagawa et al., 2019; Tong et al., 2020; Yang et al., 2020). Specific to programming, Wang et al. (2017) used a sequence of embedded program submissions to train RNNs to predict performance in the

current or the next programming exercise. However as pointed out in Abdelrahman et al. (2022), handling of non-textual content as in images, mathematical equations or code snippets to learn richer embedding representations of questions or knowledge concepts remains relatively unexplored in the domain of knowledge tracing.

Grading and feedback. While technological developments have made it easier to provide content to learners at scale, scoring their submitted work and providing feedback on similar scales remains a difficult problem. While assessing multiple-choice and fill-in-the-blank type questions is easy enough to automate, automating assessment of open-ended questions (e.g., short answers, essays, reports, code samples) and questions requiring multi-step reasoning (e.g., theorem proving, mathematical derivations) is equally hard. But automatic evaluation remains an important problem not only because it reduces the burden on teaching assistants and graders, but also removes grader-to-grader variability in assessment and helps accelerate the learning process for students by providing real-time feedback (Srikant and Aggarwal, 2014).

In the context of written prose, a number of Automatic Essay Scoring (AES) and Automatic Short Answer Grading (ASAG) systems have been developed to reliably evaluate compositions produced by learners in response to a given prompt, and are typically trained on a large set of written samples pre-scored by expert raters (Shermis and Burstein, 2003; Dikli, 2006). Over the last decade, AI-based essay grading tools evolved from using handcrafted features such as word/sentence count, mean word/sentence length, n-grams, word error rates, POS tags, grammar, and punctuation (Adamson et al., 2014; Phandi et al., 2015; Cummins et al., 2016; Contreras et al., 2018) to automatically extracted features using deep neural network variants (Taghipour and Ng, 2016; Dasgupta et al., 2018; Nadeem et al., 2019; Uto and Okano, 2020). Such systems have been developed not only to provide holistic scoring (assessing essay quality with a single score), but also for more fine-grained evaluation by providing scoring along specific dimensions of essay quality, such as organization (Persing et al., 2010), prompt-adherence (Persing and Ng, 2014), thesis clarity (Persing and Ng, 2013), argument strength (Persing and Ng, 2015), and thesis strength (Ke et al., 2019). Since it is often expensive to obtain expert-rated essays to train on each time a new prompt is introduced, considerable attention has been given to cross-prompt scoring using multi-task, domain adaptation, or transfer learning techniques, both with handcrafted (Phandi et al., 2015; Cummins et al., 2016) and automatically extracted features (Li et al., 2020; Song et al., 2020). Moreover, feedback being a critical aspect of essay drafting and revising, AES systems are increasingly being adopted into Automated Writing Evaluation (AWE) systems that provide formative feedback along with (or instead of) final scores and therefore have greater pedagogical usefulness (Hockly, 2019). For example, AWE systems have been developed for providing feedback on errors in grammar, usage and mechanics (Burstein et al., 2004) and text evidence usage in response-to-text student writings (Zhang H. et al., 2019).

AI-based evaluation tools are also heavily used in computer science education, particularly programming, due to its inherent structure and logic. Traditional approaches for automated grading of source codes such as test-case based assessments (Douce et al., 2005) and assessments using code metrics (e.g., lines of code, number of variables, number of statements), while simple, are neither robust nor effective at evaluating program quality.

A more useful direction measures similarities between abstract representations (control flow graphs, system dependence graphs) of the student's program and correct implementations of the program (Wang et al., 2007; Vujošević-Janičić et al., 2013) for automatic grading. Such similarity measurements could also be used to construct meaningful clusters of source codes and propagate feedback on student submissions based on the cluster they belong to Huang et al. (2013); Mokbel et al. (2013). Srikant and Aggarwal (2014) extracts informative features from abstract representations of the code to train machine learning models using expert-rated evaluations in order to output a finer-grained evaluation of code quality. Piech et al. (2015b) used RNNs to learn program embeddings that can be used to propagate human comments on student programs to orders of magnitude more submissions. A bottleneck in automatic program evaluation is the availability of labeled code samples. Approaches proposed to overcome this issue include learning question-independent features from code samples (Singh et al., 2016; Tarcsay et al., 2022) or zero-shot learning using human-in-the-loop rubric sampling (Wu et al., 2019).

Elsewhere, driven by the maturing of automatic speech recognition technology, AI-based assessment tools have been used for mispronunciation detection in computer-assisted language learning (Li et al., 2009, 2016; Zhang et al., 2020) or the more complex problem of spontaneous speech evaluation where the student's response is not known *apriori* (Shashidhar et al., 2015). Mathematical language processing (MLP) has been used for automatic assessment of open response mathematical questions (Lan et al., 2015; Baral et al., 2021), mathematical derivations (Tan et al., 2017), and geometric theorem proving (Mendis et al., 2017), where grades for previously unseen student solutions are predicted (or propagated from expert-provided grades), sometimes along with partial credit assignment. Zhang et al. (2022), moreover, overcomes the limitation of having to train a separate model per question by using multi-task and meta-learning tools that promote generalizability to previously unseen questions.

Academic integrity issues. Another aspect of performance assessment and monitoring is to ensure the upholding of academic integrity by detecting plagiarism and other forms of academic or research misconduct. Foltýnek et al. (2019) in their review paper on academic plagiarism detection in text (e.g., essays, reports, research papers) classifies plagiarism forms according to an increasing order of obfuscation level, from verbatim and near-verbatim copying to translation, paraphrasing, idea-preserving plagiarism, and ghostwriting. In a similar fashion, plagiarism detection methods have been developed for increasingly complex types of plagiarism, and widely adopt NLP and ML-based techniques for each (Foltýnek et al., 2019). For example, lexical detection methods use n-grams (Alzahrani, 2015) or vector space models (Vani and Gupta, 2014) to create document representations that are subsequently thresholded or clustered (Vani and Gupta, 2014) to identify suspicious documents. Syntax-based methods rely on Part-of-speech (PoS) tagging (Gupta et al., 2014), frequency of PoS tags (Hürlimann et al., 2015), or comparison of syntactic trees (Tschuggnall and Specht, 2013). Semantics-based methods

employ techniques such as word embeddings (Ferrero et al., 2017), Latent Semantic Analysis (Soleman and Purwarianti, 2014), Explicit Semantic Analysis (Meuschke et al., 2017), and word alignment (Sultan et al., 2014), often in conjunction with other ML-based techniques for downstream classification (Alfikri and Purwarianti, 2014; Hänig et al., 2015). Complementary to such textual analysis-based methods, approaches that use non-textual elements like citations, math expressions, figures, etc. also adopt machine learning for plagiarism detection (Pertile et al., 2016). Foltýnek et al. (2019) also provides a comprehensive summary of how classical ML algorithms such as tree-based methods, SVMs and neural networks have been successfully used to combine more than one type of detection method to create the best-performing meta-system. More recently, deep learning models such as different variants of convolutional and recurrent neural network architectures have also been used for plagiarism detection (El Mostafa Hambi, 2020; El-Rashidy et al., 2022).

In computer science education where programming assignments are given to evaluate students, source code plagiarism can also been classified based on increasing levels of obfuscation (Faidhi and Robinson, 1987). The detection process typically involves transforming the code into a high-dimensional feature representation followed by measurement of code similarity. Aside from traditionally used features extracted based on structural or syntactic properties of programs (Ji et al., 2007; Lange and Mancoridis, 2007), NLP-based approaches such as n-grams (Ohmann and Rahal, 2015), topic modeling (Ullah et al., 2021), character and word embeddings (Manahi, 2021), and character-level language models (Katta, 2018) are increasingly being used for robust code representations. Similarly for downstream similarity modeling or classification, unsupervised (Acampora and Cosma, 2015) and supervised (Bandara and Wijayarathna, 2011; Manahi, 2021) machine learning and deep learning algorithms are popularly used.

It is worth noting that AI itself makes plagiarism detection an uphill battle. With the increasing prevalence of easily accessible large language models like InstructGPT (Ouyang L. et al., 2022) and ChatGPT (Blog, 2022) that are capable of producing natural-sounding essays and short answers, and even working code snippets in response to a text prompt, it is now easier than ever for dishonest learners to misuse such systems for authoring assignments, projects, research papers or online exams. How plagiarism detection approaches, along with teaching and evaluation strategies, evolve around such systems remains to be seen.

(ii) **Teacher-focused**: Teaching Quality Evaluations (TQEs) are important sources of information in determining teaching effectiveness and in ensuring learning objectives are being met. The findings can be used to improve teaching skills through appropriate training and support, and also play a significant role in employment and tenure decisions and the professional growth of teachers. Such evaluations have been traditionally performed by analyzing student evaluations, teacher mutual evaluations, teacher self-evaluations and expert evaluations (Hu, 2021), which are labor-intensive to analyze at scale. Machine learning and deep learning algorithms can help with teacher evaluation by performing sentiment analysis of student comments on teacher performance (Esparza et al., 2017; Gutiérrez et al., 2018; Onan, 2020), which provides a snapshot of student attitudes toward teachers and their overall learning experiences. Further, such quantified sentiments and emotional valence scores have been used to predict students' recommendation scores for teachers in order to determine prominent factors that influence student evaluations (Okoye et al., 2022). Vijayalakshmi et al. (2020) uses student ratings related to class planning, presentation, management, and student participation to directly predict instructor performance.

Apart from helping extract insights from teacher evaluations, AI can also be used to evaluate teaching strategies on the basis of other data points from the learning process. For example, Duzhin and Gustafsson (2018) used a symbolic regression-based approach to evaluate the impact of assignment structures and collaboration type on student scores, which course instructors can use for the purpose of self-evaluation. Several works use a combination of student ratings and attributes related to the course and the instructor to predict instructor performance and investigate factors affecting learning outcomes (Mardikyan and Badur, 2011; Ahmed et al., 2016; Abunasser et al., 2022) .

4.3.3. Outcome prediction

While a course is ongoing, one way to assess knowledge development in students is through graded assignments and projects. On the other hand, educators can also benefit from automatic prediction of students' performance and automatic identification of students at risk of course non-completion. This can be accomplished by monitoring students' patterns of engagement with the course material in association with their demographic information. Such *apriori* understanding of a student's outcome allows for designing effective intervention strategies. Presently, most K-12, undergraduate and graduate students, when necessary resources are available, rely on computer and web-based infrastructure (Bulman and Fairlie, 2016). A rich source of data indicating student state is therefore generated when a student interacts with the course modules. Prior to computers being such an integral component in education, researchers frequently used surveys and questionnaires to gauge student engagement, sentiment, and attrition probability. In this section we will summarize research developments in the field of AI that *generate early prediction of student outcomes—both final performance and possibility of drop-out*.

Early research in outcome prediction focused on building explanatory regression-based models for understanding student retention using college records (Dey and Astin, 1993). The active research direction in this space gradually shifted to tackling the more complex and more actionable problems of understanding whether a student will complete a program (Dekker et al., 2009), estimating the time a student will take to complete a degree (Herzog, 2006) and predicting the final performance of a student (Nghe et al., 2007) given the current student state. In the subsequent paragraphs, we will be discussing the research contributions for outcome prediction with distinction between performance prediction in assessments and course attrition prediction. Note that

we discuss these separately as poor performance in any assessment cannot be generalized into a course non-completion.

(i) ***Apriori* performance prediction**: *Apriori* prediction of performance of a student has several benefits—it allows a student to evaluate their course selection, and allows educators to evaluate progress and offer additional assistance as needed. Not surprisingly therefore AI-based methods have been proposed to automate this important task in the education process.

Initial research articles predicting performance estimated time to degree completion (Herzog, 2006) using student demographic, academic, residential and financial aid information, student parent data and school transfer records. In a related theme, researchers have also mapped the question of performance prediction into a final exam grade prediction problem (e.g., excellent, good, fair, fail; Nghe et al., 2007; Bydžovská, 2016; Dien et al., 2020). This granular prediction eventually allows educators to assess which students require additional tutoring. Baseline algorithms in this context are Decision Trees, Support Vector Machines, Random Forests, Artificial Neural Networks etc. (regression or classification based on the problem setup). Researchers have aimed to improve the performance of the predictors by including relevant information such as student engagement, interactions (Ramesh et al., 2013; Bydžovská, 2016), role of external incentives (Jiang et al., 2014), and previous performance records (Tamhane et al., 2014). Xu et al. (2017) proposed that a student's performance or when the student anticipates graduation should be predicted progressively (using an ensemble machine learning method) over the duration of the student's tenure as the academic state of the student is ever-evolving and can be traced through their student records. The process of generalizing performance prediction to non-traditional modes of learning such as hybrid or blended learning and on-line learning has benefitted from the inclusion of additional information sources such as web-browsing information (Trakunphutthirak et al., 2019), discussion forum activity and student study habits (Gitinabard et al., 2019).

In addition to exploring a more informative and robust feature set, recently, deep learning based approaches have been identified to outperform traditional machine learning algorithms. For example, Waheed et al. (2020) used deep feed-forward neural networks and split the problem of predicting student grade into multiple binary classification problems viz., Pass-Fail, Distinction-Pass, Distinction-Fail, Withdrawn-Pass. Tsiakmaki et al. (2020) analyzed if transfer learning (i.e., pre-training neural networks on student data on a different course) can be used to accurately predict student performance. Chui et al. (2020) used a generative adversarial network based architecture, to address the challenges of low volume of training data in alternative learning paradigms such as supportive learning. Dien et al. (2020) proposed extensive data pre-processing using min-max scaler, quantile transformation, etc. before passing the data in a deep-learning model such as one-dimensional convolutional network (CN1D) or recurrent neural networks. For a comprehensive survey of ML approaches for this topic, we would refer readers to Rastrollo-Guerrero et al. (2020) and Hellas et al. (2018).

(ii) ***Apriori* attrition prediction**: Students dropping out before course completion is a concerning trend. This is more so in developing nations where very few students finish primary school (Knofczynski, 2017). The outbreak of the COVID-19 pandemic exacerbated the scenario due to indefinite school closures. This led to loss in learning and progress toward providing access to quality education (Moscoviz and Evans, 2022). The causes for dropping out of a course or a degree program can be diverse, but early prediction of it allows administrative staff and educators to intervene. To this end, there have been efforts in using machine learning algorithms to predict attrition.

Massive Open Online Courses (MOOCs): In the context of attrition, special mention must be made of Massive Open Online Courses (MOOCs). While MOOCs promise the democratization of education, one of the biggest concerns with MOOCs is the disparity between the number of students who sign up for a course versus the number of students who actually complete the course—the drop-out rate in MOOCs is significantly high (Hollands and Kazi, 2018; Reich and Ruipérez-Valiente, 2019). Yet in order to make post-secondary and professional education more accessible, MOOCs have become more a practical necessity than an experiment. The COVID-19 pandemic has only emphasized this necessity (Purkayastha and Sinha, 2021). In our literature search phase, we found a sizeable number of contributions in attrition prediction that uses data from MOOC platforms. In this subsection, we will be including those as well as attrition prediction in traditional learning environments.

Early educational data mining methods (Dekker et al., 2009) proposed to predict student drop-out mostly used data sources such as student records (i.e., student demographics, academic, residential, gap year, financial aid information) and administrative records (major administrative changes in education, records of student transfers) to train simple classifiers such as Logistic Regression, Decision Tree, BayesNet, and Random Forest. Selecting an appropriate set of features and designing explainable models has been important as these later inform intervention (Aguiar et al., 2015). To this end, researchers have explored features such as students' prior experiences, motivation and home environment (DeBoer et al., 2013) and student engagement with the course (Aguiar et al., 2014; Ramesh et al., 2014). With the inclusion of an online learning component (particularly relevant for MOOCs), click-stream data and browser information generated allowed researchers to better understand student behavior in an ongoing course. Using historical click-stream data in conjuction with present click-stream data, allowed (Kloft et al., 2014) to effectively predict drop-outs weekly using a simple Support Vector Machine algorithm. This kind of data has also been helpful in understanding the traits indicative of decreased engagement (Sinha et al., 2014), the role of a social cohort structure (Yang et al., 2013) and the sentiment in the student discussion boards and communities (Wen et al., 2014) leading up to student drop-out. He et al. (2015) addresses the concern that weekly prediction of probability of a student dropping out might have wide variance by including smoothing techniques. On the other hand, as resources to intervene might be limited, Lakkaraju et al. (2015) recommends assigning a risk-score per student rather than a binary label. Brooks et al. (2015) considers the level of activity of a student in bins of time during a semester as a binary features (active vs. inactive) and then uses these sequences as n-grams to predict drop-out. Recent developments in

predicting student attrition propose the use of data acquired from disparate sources in addition to more sophisticated algorithms such as deep feed-forward neural networks (Imran et al., 2019) and hybrid logit leaf model (Coussement et al., 2020).

5. Discussion

In this article, we have investigated the involvement of artificial intelligence in the end-to-end educational process. We have highlighted specific research problems both in the planning and in the knowledge delivery phase and reviewed the technological progress in addressing those problems in the past two decades. To the best of our knowledge, such distinction between proactive and reactive phases of education accompanied by a technical deep-dive is an uniqueness of this review.

5.1. Major trends in involvement of AI in the end-to-end education process

The growing interest in AIEd can be inferred from Figures 2, 4 which show how both the count of technical contributions and the count of review articles on the topic have increased over the past two decades. It is to be noted that the number of technical contributions in 2021 and 2022 (assuming our sample of reviewed articles is representative of the population) might have fallen in part due to pandemic-related indefinite school closures and shift to alternate learning models. This triggered a setback on data collection, reporting, and annotation efforts due to a number of factors including lack of direct access to participants, unreliable network connectivity and the necessity of enumerators adopting to new training modes (Wolf et al., 2022). Another important observation from Figure 3 is that AIEd research in most categories focuses heavily on learners in universities, e-learning platforms and MOOCs—work targeting pre-school and K-12 learners is conspicuously absent. A notable exception is research surrounding tutoring aids that has a nearly uniform attention for different target audience groups.

In all categories, to different extents, we see a distinct shift from rule-based and statistical approaches to classical ML to deep learning methods, and from handcrafted features to automatically extracted features. This advancement goes hand-in-hand with the increasingly complex nature of the data being utilized for training AIEd systems. Whereas, earlier approaches used mostly static data (e.g., student records, administrative records, demographic information, surveys, and questionnaires), the use of more sophisticated algorithms necessitated (and in turn benefited from) more real-time and high-volume data (e.g., student-teacher/peer-peer interaction data, click-stream information, web-browsing data). The type of data used by AIEd systems also evolved from mostly tabular records to more text-based and even multi-modal data, spurred on by the emergence of large language models that can handle large quantities of such data.

Even though data-hungry models like deep neural networks have grown in popularity across almost all categories discussed here, AIEd often suffers from the availability of sufficient labeled data to train such systems. This is particularly true for small classes and new course offerings, or when existing curriculum or tests are changed to incorporate new elements. As a result, another emerging trend in AIEd focuses on using information from resource-rich courses or existing teaching/evaluation content through domain adaptation, transfer learning, few-shot learning, meta learning, etc.

5.2. Impact of COVID-19 pandemic on driving AI research in the frontier of education

COVID-19 pandemic, possibly the most significant social disruptor in recent history, impacted more than 1.5 billion students worldwide (UNESCO, 2022) and is believed to have had far-reaching consequences in the domain of education, possibly even generational setbacks (Tadesse and Muluye, 2020; Dorn et al., 2021; Spector, 2022). As lockdowns and social distancing mandated a hastened transition to fully virtual delivery of educational content, the pandemic era saw an increasing adoption of video conferencing softwares and social media platforms for knowledge delivery, combined with more asynchronous formats of learning. These alternative media of communication were often accompanied by decreasing levels of engagement and satisfaction of learners (Wester et al., 2021; Hollister et al., 2022). There was also a corresponding decrease in practical sessions, labs, and workshops, which are quite critical in some fields of education (Hilburg et al., 2020). However, the pandemic also led to an accelerated adoption of AI-based approaches in education. Pilot studies show that the pandemic led to a significant increase in the usage of AI-based e-learning platforms (Pantelimon et al., 2021). Moreover, a natural by-product of the transition to online learning environments is the generation and logging of more data points from the learning process (Xie et al., 2020) that can be used in AI-based methods to assess and drive student engagement and provide personalized feedback. Online teaching platforms also make it easier to incorporate web-based content, smart interactive elements and asynchronous review sessions to keep students more engaged (Kexin et al., 2020; Pantelimon et al., 2021).

Several recent works have investigated the role of pandemic-driven remote and hybrid instruction in widening gaps in educational achievements by race, poverty level, and gender (Halloran et al., 2021; UNESCO, 2021; Goldhaber et al., 2022). A widespread transition to remote learning necessitates access to proper infrastructure (electricity, internet connectivity, and smart electronic devices that can support video conferencing apps and basic file sharing) as well as resources (learning material, textbooks, educational softwares, etc.), which create barriers for low-income groups (Muñoz-Najar et al., 2021). Even within similar populations, unequal distribution of household chores, income-generating activities, and access to technology-enabled devices affect students of different genders disproportionately (UNESCO, 2021). Moreover, remote learning requires a level of tech-savviness on the part of students and teachers alike, which might be less prevalent in people with learning disabilities. In this context, Garg and Sharma (2020) outlines the different ways AI is used in special need education for development of adaptive and inclusive pedagogies. Salas-Pilco et al. (2022) reviews the different ways

in which AI positively impacts education of minority students, e.g., through facilitating performance/engagement improvement, student retention, student interest in STEM/STEAM fields, etc. Salas-Pilco et al. (2022) also outlines the technological, pedagogical, and socio-cultural barriers for AIEd in inclusive education.

5.3. Existing challenges in adopting artificial intelligence for education

In 2023, artificial intelligence has permeated the lives of people in some aspect or other globally (e.g. chat-bots for customer service, automated credit score analysis, personalized recommendations). At the same time, AI-driven technology for the education sector is gradually becoming a practical necessity globally. The question therefore is, what are the existing barriers in global adoption of AI for education in a safe and inclusive manner—we discuss some of our observations with regards to deploying existing AI driven educational technology at scale.

5.3.1. Lack of concrete legal and ethical guidelines for AIEd research

As pointed out by Pedro et al. (2019), besides most AIEd researchers being concentrated in the technologically advanced parts of the world, most AIEd platforms and applications are owned currently by the private sector. The private investor funded research in big corporations such as Coursera, EdX, IBM, McGraw-Hill, and start-ups like Elsa, Century, Querium have yielded several robust AIEd applications. However, as these platforms are privately owned, there is little transparency and regulations regarding their development and operations. Due to this, there is growing concern on the part of guardians and teaching staff regarding the data accessed by these platforms, privacy, and security of the data stored and explainability of the deployed models. To alleviate this, regulation policies at the international, national, and state levels can help address the concerns of the end users. While many tech-savvy nations have had a head start in this Stirling et al. (2017), drafting general guidelines for AIEd platforms is still very much a nascent concept for most policy makers.

5.3.2. Lack of equitable access to infrastructure hosting AIEd

Education is one of the most important social equalizers (Winthrop, 2018). However, in order to ensure more people have access to quality education, AI-enabled teaching, and studying tools are necessary to reduce the stress on educators and administrative staff (Pedro et al., 2019). The paradox here is that the cost of deploying and operating AIEd tools often alienates communities with limited means thereby widening the gap in access to education. Nye (2015) mentions that access to electricity, internet, data storage, and processing hardware have been barriers in deploying AI-driven platforms. To remove these obstacles, changes must be brought about in local and global levels. While formation of international alliances that invest in infrastructure development can usher in the technology in developing nations, changes in local policies can expedite the process (Mbangula, 2022).

5.3.3. Lack of skilled personnels to operate AIEd tools in production

Investing in AIEd research and supporting infrastructure alone is not sufficient to ensure long term utility and usage of AI-driven tools for education. Workforce responsible for using these tools on a day-to-day basis must also be brought up to speed. Currently, there is a considerable amount of apprehension, particularly in developing countries, regarding use of AI for education (Shum and Luckin, 2019; Alam, 2021). The main concerns are related to data privacy and security, job security, ethics etc. post adoption of AI in this sector. These concerns in turn have slowed down integration of technology for education. In this context, we must echo (Pedro et al., 2019) in mentioning that while these concerns are relevant and must be addressed, in our review of AIEd research, we have not found any evidence that should invoke consternation in educators and administrative staff. AIEd research as it stands today only augments the role of the teacher, and does not eliminate it. Furthermore, for the foreseeable future, we would need a human in the loop to provide feedback and ensure proper daily usage of these tools.

5.4. Concluding remarks

Through this review, we identified the paradigm shift over the past 20 years in formulating computational models (i.e., choice of algorithms, choice of features etc.) and training them (i.e., choice of data)—we are indeed increasingly leaning toward *sophisticated yet explainable* frameworks. As the scope of this review includes a period of social disruption due to COVID-19 pandemic, it provided us the opportunity to introspect on the utility and the robustness of the proposed technology thus far. To this end, we have discussed the concerns and limitations brought to light by the pandemic and research ideas spawning from that.

With the target of ensuring equitable access to education being set for 2030 by UNGA (United Nations, 2015), one of the inevitable questions arising is: *are we ready to use AI driven ed-tech tools to support educators and students?*. This remains however a question to be answered. Based on our survey, we have observed that while in some parts of the world we have seen great momentum in making AIEd a part and parcel of the education sector, in other parts of the world this progress is stymied by inadequate access to necessary infrastructure and human resources. The ethical and legal implications for large-scale adoption of AI for education is also a topic of active debate (Holmes and Porayska-Pomsta, 2022). The pivotal point at this time is that while there needs to be changes at a socio-economic level to adopt the state of the art AI-driven ed-tech technologies as standard tools for education, the progress made and the ongoing conversations are reasons for positivity.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

A preprint version of this paper is available at: https://arxiv.org/abs/2301.10231 (Mallik and Gangopadhyay, 2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1151391/full#supplementary-material

Supplementary section contains the full list of 195 technical articles that have been reviewed in this paper under their respective categories and subcategories.

References

Abdelrahman, G., and Wang, Q. (2019). "Knowledge tracing with sequential key-value memory networks," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris), 175–184. doi: 10.1145/3331184.3331195

Abdelrahman, G., Wang, Q., and Nunes, B. P. (2022). Knowledge tracing: a survey. *ACM Comput. Surveys* 55, 1–37. doi: 10.1145/3569576

Abunasser, B. S., AL-Hiealy, M. R. J., Barhoom, A. M., Almasri, A. R., and Abu-Naser, S. S. (2022). Prediction of instructor performance using machine and deep learning techniques. *Int. J. Adv. Comput. Sci. Appl.* 13, 78–83. doi: 10.14569/IJACSA.2022.0130711

Acampora, G., and Cosma, G. (2015). "A fuzzy-based approach to programming language independent source-code plagiarism detection," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (Istanbul), 1–8. doi: 10.1109/FUZZ-IEEE.2015.7337935

Adamson, A., Lamb, A., and December, R. (2014). *Automated Essay Grading.*

Afzal, S., Dhamecha, T. I., Gagnon, P., Nayak, A., Shah, A., Carlstedt-Duke, J., et al. (2020). "AI medical school tutor: modelling and implementation," in *International Conference on Artificial Intelligence in Medicine* (Minneapolis, MN: Springer), 133–145. doi: 10.1007/978-3-030-59137-3_13

Agarwal, M., and Mannem, P. (2011). "Automatic gap-fill question generation from text books," in *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications* (Portland, OR), 56–64.

Agarwal, S. (2020). *Trade-offs between fairness, interpretability, and privacy in machine learning* (Master's thesis). University of Waterloo, Waterloo, ON, Canada.

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., and Goodrich, V. (2014). "Engagement vs. performance: using electronic portfolios to predict first semester engineering student retention," in *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (Indianapolis, IN), 103–112. doi: 10.1145/2567574.2567583

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., and Addison, K. L. (2015). "Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (Poughkeepsie, NY), 93–102. doi: 10.1145/2723576.2723619

Ahmad, K., Qadir, J., Al-Fuqaha, A., Iqbal, W., El-Hassan, A., Benhaddou, D., et al. (2020). *Data-Driven Artificial Intelligence in Education: A Comprehensive Review*. EdArXiv.

Ahmad, S. F., Alam, M. M., Rahmat, M. K., Mubarik, M. S., and Hyder, S. I. (2022). Academic and administrative role of artificial intelligence in education. *Sustainability* 14, 1101. doi: 10.3390/su14031101

Ahmed, A. M., Rizaner, A., and Ulusoy, A. H. (2016). Using data mining to predict instructor performance. *Proc. Comput. Sci.* 102, 137–142. doi: 10.1016/j.procs.2016.09.380

Ahmed, U. Z., Gulwani, S., and Karkare, A. (2013). "Automatically generating problems and solutions for natural deduction," in *Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing).

Ahn, J.-w., Tejwani, R., Sundararajan, S., Sipolins, A., O'Hara, S., Paul, A., et al. (2018). "Intelligent virtual reality tutoring system supporting open educational resource access," in *International Conference on Intelligent Tutoring Systems* (Montreal: Springer), 280–286. doi: 10.1007/978-3-319-91464-0_28

Alam, A. (2021). "Possibilities and apprehensions in the landscape of artificial intelligence in education," in *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (Nagpur), 1–8. doi: 10.1109/ICCICA52458.2021.9697272

Alfaro, L., Rivera, C., Luna-Urquizo, J., Castañeda, E., and Fialho, F. (2018). Online learning styles identification model, based on the analysis of user interactions within an e-learning platforms, using neural networks and fuzzy logic. *Int. J. Eng. Technol.* 7, 76. doi: 10.14419/ijet.v7i3.13.16328

Alfikri, Z. F., and Purwarianti, A. (2014). Detailed analysis of extrinsic plagiarism detection system using machine learning approach (naive Bayes and SVM). *TELKOMNIKA Indones. J. Electr. Eng.* 12, 7884–7894. doi: 10.11591/telkomnika.v12i11.6652

AlGhamdi, A., Barsheed, A., AlMshjary, H., and AlGhamdi, H. (2020). "A machine learning approach for graduate admission prediction," in *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing* (Singapore), 155–158. doi: 10.1145/3388818.3393716

Alkhatlan, A., and Kalita, J. (2018). Intelligent tutoring systems: a comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628*. doi: 10.5120/ijca2019918451

AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2021). "A systematic review of data-driven approaches to item difficulty prediction," in *International Conference on Artificial Intelligence in Education* (Utrecht: Springer), 29–41. doi: 10.1007/978-3-030-78292-4_3

Alzahrani, S. (2015). "Arabic plagiarism detection using word correlation in n-grams with k-overlapping approach," in *Proceedings of the Workshops at the 7th Forum for Information Retrieval Evaluation (FIRE)* (Gandhinagar), 123–125.

Arroyo, I., Beal, C., Murray, T., Walles, R., and Woolf, B. (2004). "Wayang outpost: intelligent tutoring for high stakes achievement tests," in *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS2004)* (Maceió), 468–477. doi: 10.1007/978-3-540-30139-4_44

Arroyo, I., Cooper, D. G., Burleson, W., and Woolf, B. P. (2010). "Bayesian networks and linear regression models of students–goals, moods, and emotions," in *Handbook of Educational Data Mining* eds Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. J. d. (Chapman & Hall), 323–338.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., and Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *Int. J. Artif. Intell. Educ.* 24, 387–426. doi: 10.1007/s40593-014-0023-y

Ashfaque, M. W., Tharewal, S., Iqhbal, S., and Kayte, C. N. (2020). "A review on techniques, characteristics and approaches of an intelligent tutoring chatbot system," in *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)* (Aurangabad), 258–262. doi: 10.1109/ICSIDEMPC49020.2020.9299583

Assiri, B., Bashraheel, M., and Alsuri, A. (2022). "Improve the accuracy of students admission at universities using machine learning techniques," in *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (Riyadh), 127–132. doi: 10.1109/CDMA54072.2022.00026

Baidoo-Anu, D., and Owusu Ansah, L. (2023). *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*. doi: 10.2139/ssrn.4337484

Bajaj, R., and Sharma, V. (2018). Smart education with artificial intelligence based determination of learning styles. *Proc. Comput. Sci.* 132, 834–842. doi: 10.1016/j.procs.2018.05.095

Ball, R., Duhadway, L., Feuz, K., Jensen, J., Rague, B., and Weidman, D. (2019). "Applying machine learning to improve curriculum design," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN), 787–793. doi: 10.1145/3287324.3287430

Bandara, U., and Wijayarathna, G. (2011). A machine learning based tool for source code plagiarism detection. *Int. J. Mach. Learn. Comput.* 1, 337. doi: 10.7763/IJMLC.2011.V1.50

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., and Heffernan, N. T. (2021). *Improving Automated Scoring of Student Open Responses in Mathematics*. Paris: International Educational Data Mining Society.

Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. (2020a). "Introducing a framework to assess newly created questions with natural language processing," in *International Conference on Artificial Intelligence in Education* (Ifrane: Springer), 43–54. doi: 10.1007/978-3-030-52237-7_4

Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. (2020b). "R2de: a NLP approach to estimating irt parameters of newly generated questions," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt), 412–421. doi: 10.1145/3375462.3375517

Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., et al. (2022). A survey on recent approaches to question difficulty estimation from text. *ACM Comput. Surveys* 55, 1–37. doi: 10.1145/3556538

Bernard, J., Chang, T.-W., Popescu, E., and Graf, S. (2015). "Using artificial neural networks to identify learning styles," in *International Conference on Artificial Intelligence in Education* (Madrid: Springer), 541–544. doi: 10.1007/978-3-319-19773-9_57

Blog, O. (2022). *Chatgpt: Optimizing Language Models for Dialogue*.

Bloom, B. S. (1984). The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* 13, 4–16. doi: 10.3102/0013189X013006004

Bogina, V., Hartman, A., Kuflik, T., and Shulner-Tal, A. (2022). Educating software and ai stakeholders about algorithmic fairness, accountability, transparency and ethics. *Int. J. Artif. Intell. Educ.* 32, 808–833. doi: 10.1007/s40593-021-00248-0

Botelho, A. F., Baker, R. S., and Heffernan, N. T. (2017). "Improving sensor-free affect detection using deep learning," in *International Conference on Artificial Intelligence in Education* (Wuhan: Springer), 40–51. doi: 10.1007/978-3-319-61425-0_4

Brooks, C., Thompson, C., and Teasley, S. (2015). "A time series interaction analysis method for building predictive models of learners using log data," in *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (Poughkeepsie, NY), 126–135. doi: 10.1145/2723576.2723581

Bruggink, T. H., and Gambhir, V. (1996). Statistical models for college admission and enrollment: a case study for a selective liberal arts college. *Res. High. Educ.* 37, 221–240. doi: 10.1007/BF01730116

Bryant, J., Heitz, C., Sanghvi, S., and Wagle, D. (2020). *How Artificial Intelligence Will Impact K-12 Teachers*. McKinsey.

Bulman, G., and Fairlie, R. W. (2016). "Technology and education: computers, software, and the internet," in *Handbook of the Economics of Education, Vol. 5* eds Hanushek, E. A., Machin, S., and Woessmann, L. (Elsevier), 239–280. doi: 10.1016/B978-0-444-63459-7.00005-1

Burkacky, O., Dragon, J., and Lehmann, N. (2022). *The Semiconductor Decade: A Trillion-Dollar Industry*. McKinsey.

Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: the criterion online writing service. *Ai Mag.* 25, 27. doi: 10.1609/aimag.v25i3.1774

Bydžovská, H. (2016). *A Comparative Analysis of Techniques for Predicting Student Performance*. Raleigh, NC: International Educational Data Mining Society.

Cen, H., Koedinger, K., and Junker, B. (2006). "Learning factors analysis-a general method for cognitive model evaluation and improvement," in *International Conference on Intelligent Tutoring Systems* (Jhongli: Springer), 164–175. doi: 10.1007/11774303_17

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. (2018). Artificial intelligence trends in education: a narrative overview. *Proc. Comput. Sci.* 136, 16–24. doi: 10.1016/j.procs.2018.08.233

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access* 8, 75264–75278. doi: 10.1109/ACCESS.2020.2988510

Chen, S. Y., and Wang, J.-H. (2021). Individual differences and personalized learning: a review and appraisal. *Univers. Access Inform.* Soc. 20, 833–849. doi: 10.1007/s10209-020-00753-4

Chen, X., Zou, D., Xie, H., Cheng, G., and Liu, C. (2022). Two decades of artificial intelligence in education. *Educ. Technol. Soc.* 25, 28–47.

Chen, Y., Wu, L., and Zaki, M. J. (2020). "Reinforcement learning based graph-to-sequence model for natural question generation," in *International Conference on Learning Representations*.

Chui, K. T., Liu, R. W., Zhao, M., and De Pablos, P. O. (2020). Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine. *IEEE Access* 8, 86745–86752. doi: 10.1109/ACCESS.2020.2992869

Contreras, J. O., Hilles, S., and Abubakar, Z. B. (2018). "Automated essay scoring with ontology based on text mining and NLTK tools," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)* (Selangor), 1–6. doi: 10.1109/ICSCEE.2018.8538399

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User Adapt. Interact.* 4, 253–278. doi: 10.1007/BF01099821

Correia, R., Baptista, J., Eskenazi, M., and Mamede, N. (2012). "Automatic generation of cloze question stems," in *International Conference on Computational Processing of the Portuguese Language* (Coimbra: Springer), 168–178. doi: 10.1007/978-3-642-28885-2_19

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., and Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: the beneficial impact of the logit leaf model. *Decis. Support Syst.* 135, 113325. doi: 10.1016/j.dss.2020.113325

Cummins, R., Zhang, M., and Briscoe, E. (2016). *Constrained Multi-Task Learning for Automated Essay Scoring*. Association for Computational Linguistics. doi: 10.18653/v1/P16-1075

Dasgupta, T., Naskar, A., Dey, L., and Saha, R. (2018). "Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring," in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (Melbourne), 93–102. doi: 10.18653/v1/W18-3713

Deane, P., and Sheehan, K. (2003). "Automatic item generation via frame semantics: Natural language generation of math word problems," in *Annual Meeting of the National Council of Measurement in Education (ERIC)*.

DeBoer, J., Stump, G. S., Seaton, D., Ho, A., Pritchard, D. E., and Breslow, L. (2013). "Bringing student backgrounds online: MOOC user demographics, site usage, and online learning," in *Educational Data Mining 2013* (Memphis, TN).

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., et al. (2018). Detecting and addressing frustration in a serious game for military training. *Int. J. Artif. Intell. Educ.* 28, 152–193. doi: 10.1007/s40593-017-0152-1

Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). "Predicting students drop out: a case study," in *International Working Group on Educational Data Mining*.

Dey, E. L., and Astin, A. W. (1993). Statistical alternatives for studying college student retention: a comparative analysis of logit, probit, and linear regression. *Res. High. Educ.* 34, 569–581. doi: 10.1007/BF00991920

Dien, T. T., Luu, S. H., Thanh-Hai, N., and Thai-Nghe, N. (2020). Deep learning with data transformation and factor analysis for student performance prediction. *Int. J. Adv. Comput. Sci. Appl.* 11, 711–721. doi: 10.14569/IJACSA.2020.0110886

Dikli, S. (2006). An overview of automated scoring of essays. *J. Technol. Learn. Assess.* 5.

Dong, N., and Chen, Z. (2020). *The Fourth Education Revolution: Will Artificial Intelligence Liberate or Infantilise Humanity: Buckingham, University of Buckingham*. Springer.

Dorn, E., Hancock, B., Sarakatsannis, J., and Viruleg, E. (2021). *COVID-19 and Education: The Lingering Effects of Unfinished Learning*. McKinsey. Available online at: https://www.mckinsey.com/industries/education/our-insights/covid-19-and-education-the-lingering-effects-of-unfinished-learning

Doroudi, S. (2019). *Integrating human and machine intelligence for enhanced curriculum design* (Ph.D. dissertation). Pittsburgh, PA: Air Force Research Laboratory.

Douce, C., Livingstone, D., and Orwell, J. (2005). Automatic test-based assessment of programming: a review. *J. Educ. Resour. Comput.* 5, 4–es. doi: 10.1145/1163405.1163409

Dreyfus, H. L. (1999). Anonymity versus commitment: the dangers of education on the internet. *Ethics Inform. Technol.* 1, 15–20.

Du, X., Shao, J., and Cardie, C. (2017). "Learning to ask: neural question generation for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Vancouver), 1342–1352. doi: 10.18653/v1/P17-1123

Duzhin, F., and Gustafsson, A. (2018). Machine learning-based app for self-evaluation of teacher-specific instructional style and tools. *Educ. Sci.* 8:7. doi: 10.3390/educsci8010007

El Guabassi, I., Bousalem, Z., Marah, R., and Qazdar, A. (2021). A recommender system for predicting students' admission to a graduate program using machine learning algorithms. *Int. J. Online Biomed. Engg.* 17, 135–147. doi: 10.3991/ijoe.v17i02.20049

El Mostafa Hambi, F. B. (2020). A new online plagiarism detection system based on deep learning. *Int. J. Adv. Comput. Sci. Appl.* 11, 470–478. doi: 10.14569/IJACSA.2020.0110956

El-Rashidy, M. A., Mohamed, R. G., El-Fishawy, N. A., and Shouman, M. A. (2022). Reliable plagiarism detection system based on deep learning approaches. *Neural Comput. Appl.* 34, 18837–18858. doi: 10.1007/s00521-022-07486-w

Emelianov, V., Gast, N., Gummadi, K. P., and Loiseau, P. (2020). "On fair selection in the presence of implicit variance," in *Proceedings of the 21st ACM Conference on Economics and Computation* (Hungary), 649–675. doi: 10.1145/3391403.3399482

Esparza, G. G., de Luna, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Álvarez, M., et al. (2017). "A sentiment analysis model to analyze students reviews of teacher performance using support vector machines," in *International Symposium on Distributed Computing and Artificial Intelligence* (Porto: Springer), 157–164. doi: 10.1007/978-3-319-62410-5_19

Fahimirad, M., and Kotamjani, S. S. (2018). A review on application of artificial intelligence in teaching and learning in educational contexts. *Int. J. Learn. Dev.* 8, 106–118. doi: 10.5296/ijld.v8i4.14057

Faidhi, J. A., and Robinson, S. K. (1987). An empirical approach for detecting program similarity and plagiarism within a university programming environment. *Comput. Educ.* 11, 11–19. doi: 10.1016/0360-1315(87)90042-X

Fang, J., Zhao, W., and Jia, D. (2019). "Exercise difficulty prediction in online education systems," in *2019 International Conference on Data Mining Workshops (ICDMW)* (Beijing), 311–317. doi: 10.1109/ICDMW.2019.00053

Feenberg, A. (2017). The online education controversy and the future of the university. *Found. Sci.* 22, 363–371. doi: 10.1007/s10699-015-9444-9

Felder, R. M. (1988). Learning and teaching styles in engineering education. *Engg. Educ.* 78, 674–681.

Ferrero, J., Besacier, L., Schwab, D., and Agnès, F. (2017). "Using word embedding for cross-language plagiarism detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2* (Valencia), 415–421. doi: 10.18653/v1/E17-2066

Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A.-A., et al. (2021). "Bridging machine learning and mechanism design towards algorithmic fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* , 489–503. doi: 10.1145/3442188.3445912

Foltýnek, T., Meuschke, N., and Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. *ACM Comput. Surveys* 52, 1–42. doi: 10.1145/3345317

Garg, S., and Sharma, S. (2020). Impact of artificial intelligence in special need education to promote inclusive pedagogy. *Int. J. Inform. Educ. Technol.* 10, 523–527. doi: 10.18178/ijiet.2020.10.7.1418

Ghosh, A., Heffernan, N., and Lan, A. S. (2020). "Context-aware attentive knowledge tracing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (California, CA), 2330–2339. doi: 10.1145/3394486.3403282

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., and Lynch, C. F. (2019). How widely can prediction models be generalized? An analysis of performance prediction in blended courses. *IEEE Transactions on Learning Technologies*. 12, 184–197. doi: 10.1109/TLT.2019.2911832

Goh, S. L., Kendall, G., and Sabar, N. R. (2019). Simulated annealing with improved reheating and learning for the post enrolment course timetabling problem. *J. Oper. Res. Soc.* 70, 873–888. doi: 10.1080/01605682.2018.1468862

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., and Staiger, D. O. (2022). *The Consequences of Remote and Hybrid Instruction During the Pandemic*. Technical report, National Bureau of Economic Research. doi: 10.3386/w30010

Goni, M. O. F., Matin, A., Hasan, T., Siddique, M. A. I., Jyoti, O., and Hasnain, F. M. S. (2020). "Graduate admission chance prediction using deep neural network," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)* (Bhubaneswar), 259–262.

Gordon, A., van Lent, M., Van Velsen, M., Carpenter, P., and Jhala, A. (2004). "Branching storylines in virtual reality environments for leadership development," in *Proceedings of the National Conference on Artificial Intelligence* (Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press), 844–851.

Gordon, G., and Breazeal, C. (2015). "Bayesian active learning-based robot tutor for children's word-reading skills," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Austin, TX), Vol. 29. doi: 10.1609/aaai.v29i1.9376

Green, D., Walsh, T., Cohen, P., and Chang, Y.-H. (2011). "Learning a skill-teaching curriculum with dynamic Bayes nets," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 25* (San Francisco, CA), 1648–1654. doi: 10.1609/aaai.v25i2.18855

Grivokostopoulou, F., Hatzilygeroudis, I., and Perikos, I. (2014). Teaching assistance and automatic difficulty estimation in converting first order logic to clause form. *Artif. Intell. Rev.* 42, 347–367. doi: 10.1007/s10462-013-9417-8

Gupta, D., Vani, K., and Singh, C. K. (2014). "Using natural language processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (Delhi), 2694–2699. doi: 10.1109/ICACCI.2014.6968314

Gutiérrez, G., Canul-Reich, J., Zezzatti, A. O., Margain, L., and Ponce, J. (2018). Mining: students comments about teacher performance assessment using machine learning algorithms. *Int. J. Combin. Optim. Probl. Inform.* 9, 26.

Haenlein, M., and Kaplan, A. (2019). A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *Calif. Manage. Rev.* 61, 5–14. doi: 10.1177/0008125619864925

Halloran, C., Jack, R., Okun, J. C., and Oster, E. (2021). *Pandemic Schooling Mode and Student Test Scores: Evidence From US States*. Technical report, National Bureau of Economic Research. doi: 10.3386/w29497

Hänig, C., Remus, R., and De La Puente, X. (2015). "EXB themis: extensive feature extraction from word alignments for semantic textual similarity," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, TX), 264–268. doi: 10.18653/v1/S15-2046

Harley, J. M., Lajoie, S. P., Frasson, C., and Hall, N. C. (2017). Developing emotion-aware, advanced learning technologies: a taxonomy of approaches and features. *Int. J. Artif. Intell. Educ.* 27, 268–297. doi: 10.1007/s40593-016-0126-8

He, J., Bailey, J., Rubinstein, B., and Zhang, R. (2015). "Identifying at-risk students in massive open online courses," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29* (Austin, TX). doi: 10.1609/aaai.v29i1.9471

Heilman, M. (2011). *Automatic factual question generation from text* (Ph.D. thesis). Carnegie Mellon University, Pittsburgh, PA, United States.

Heilman, M., and Smith, N. A. (2010). "Good question! Statistical ranking for question generation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Los Angeles, CA), 609–617.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., et al. (2018). "Predicting academic performance: a systematic literature review," in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (Larnaca), 175–199. doi: 10.1145/3293881.3295783

Herzog, S. (2006). Estimating student retention and degree-completion time: decision trees and neural networks vis-à-vis regression. *New Direct. Instit. Res.* 131, 17–33. doi: 10.1002/ir.185

Hilburg, R., Patel, N., Ambruso, S., Biewald, M. A., and Farouk, S. S. (2020). Medical education during the coronavirus disease-2019 pandemic: learning from a distance. *Adv. Chron. Kidney Dis.* 27, 412–417. doi: 10.1053/j.ackd.2020.05.017

Hockly, N. (2019). Automated writing evaluation. *ELT. J.* 73, 82–88. doi: 10.1093/elt/ccy044

Hollands, F., and Kazi, A. (2018). *Benefits and Costs of MOOC-Based Alternative Credentials*. Center for Benefit-Cost Studies of Education.

Hollister, B., Nair, P., Hill-Lindsay, S., and Chukoskie, L. (2022). Engagement in online learning: student attitudes and behavior during COVID-19. *Front. Educ.* 7, 851019. doi: 10.3389/feduc.2022.851019

Holmes, W., and Porayska-Pomsta, K. (2022). *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*. Taylor & Francis. doi: 10.4324/9780429329067

Holmes, W., and Tuomi, I. (2022). State of the art and practice in AI in education. *Eur. J. Educ.* 57, 542–570. doi: 10.1111/ejed.12533

Honey, P., and Mumford, A. (1986). *The Manual of Learning Styles*.

Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. *Int. J. Emerg. Technol. Learn.* 16, 87–101. doi: 10.3991/ijet.v16i05.20299

Huang, J., Piech, C., Nguyen, A., and Guibas, L. (2013). "Syntactic and functional variability of a million code submissions in a machine learning MOOC," in *AIED 2013 Workshops Proceedings, Vol. 25* (Memphis, TN).

Huang, J., Saleh, S., and Liu, Y. (2021). A review on artificial intelligence in education. *Acad. J. Interdisc. Stud.* 10, 206. doi: 10.36941/ajis-2021-0077

Huang, Y., Huang, W., Tong, S., Huang, Z., Liu, Q., Chen, E., et al. (2021). "Stan: adversarial network for cross-domain question difficulty prediction," in *2021 IEEE International Conference on Data Mining (ICDM)* (Auckland), 220–229. doi: 10.1109/ICDM51629.2021.00032

Huang, Y.-T., Tseng, Y.-M., Sun, Y. S., and Chen, M. C. (2014). "Tedquiz: automatic quiz generation for ted talks video clips to assess listening comprehension," in *2014 IEEE 14Th International Conference on Advanced Learning Technologies* (Athens), 350–354. doi: 10.1109/ICALT.2014.105

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., et al. (2017). "Question difficulty prediction for reading problems in standard tests," in *Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA). doi: 10.1609/aaai.v31i1.10740

Humble, N., and Mozelius, P. (2019). "Artificial intelligence in education–a promise, a threat or a hype," in *Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics* (Oxford), 149–156.

Hürlimann, M., Weck, B., van den Berg, E., Suster, S., and Nissim, M. (2015). "Glad: Groningen lightweight authorship detection," in *CLEF (Working Notes)* (Toulouse).

Hwang, G.-J., Xie, H., Wah, B. W., and Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Comput. Educ. Artif. Intell.* 1, 10001. doi: 10.1016/j.caeai.2020.100001

Idris, N., Yusof, N., and Saad, P. (2009). Adaptive course sequencing for personalization of learning path using neural network. *Int. J. Adv. Soft Comput. Appl.* 1, 49–61.

Imran, A. S., Dalipi, F., and Kastrati, Z. (2019). "Predicting student dropout in a MOOC: an evaluation of a deep neural network model," in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence* (Bali), 190–195. doi: 10.1145/3330482.3330514

Indurthi, S. R., Raghu, D., Khapra, M. M., and Joshi, S. (2017). "Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 1* (Valencia), 376–385.

Islam, M. Z., Ali, R., Haider, A., Islam, M. Z., and Kim, H. S. (2021). Pakes: a reinforcement learning-based personalized adaptability knowledge extraction strategy for adaptive learning systems. *IEEE Access* 9, 155123–155137. doi: 10.1109/ACCESS.2021.3128578

Jackson, G. T., and McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *J. Educ. Psychol.* 105, 1036. doi: 10.1037/a0032580

Jamison, J. (2017). "Applying machine learning to predict Davidson college's admissions yield," in *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, WA), 765–766. doi: 10.1145/3017680.3022468

Ji, J.-H., Woo, G., and Cho, H.-G. (2007). "A source code linearization technique for detecting plagiarized programs," in *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* (Dundee), 73–77. doi: 10.1145/1269900.1268807

Jiang, S., Williams, A., Schenke, K., Warschauer, M., and O'dowd, D. (2014). "Predicting MOOC performance with week 1 behavior," in *Educational Data Mining 2014* (London).

Jouault, C., and Seta, K. (2013). "Building a semantic open learning space with adaptive question generation support," in *Proceedings of the 21st International Conference on Computers in Education* (Bali), 41–50.

Kalady, S., Elikkottil, A., and Das, R. (2010). "Natural language question generation using syntax and keywords," in *Proceedings of QG2010: The Third Workshop on Question Generation, Vol. 2* (Pittsburgh, PA), 5–14.

Katta, J. Y. B. (2018). *Machine learning for source-code plagiarism detection* (Ph.D. thesis). International Institute of Information Technology Hyderabad.

Ke, Z., Inamdar, H., Lin, H., and Ng, V. (2019). "Give me more feedback ii: annotating thesis strength and related attributes in student essays," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence), 3994–4004.

Keller, S. U. (2021). Automatic generation of word problems for academic education via natural language processing (nlp). *arXiv preprint arXiv:2109.13123*. doi: 10.48550/arXiv.2109.13123

Kenekayoro, P. (2019). Incorporating machine learning to evaluate solutions to the university course timetabling problem. *Covenant J. Inform. Commun. Technol.* 7:18–35. doi: 10.48550/arXiv.2010.00826

Kessels, J. (1999). "A relational approach to curriculum design," in *Design Approaches and Tools in Education and Training* (Springer), 59–70. doi: 10.1007/978-94-011-4255-7_5

Kexin, L., Yi, Q., Xiaoou, S., and Yan, L. (2020). "Future education trend learned from the COVID-19 pandemic: take artificial intelligence online course as an example," in *2020 International Conference on Artificial Intelligence and Education (ICAIE)* (Tianjin), 108–111. doi: 10.1109/ICAIE50891.2020.00032

Khajah, M., Wing, R., Lindsey, R. V., and Mozer, M. (2014). "Integrating latent-factor and knowledge-tracing models to predict individual differences in learning," in *EDM* (London), 99–106.

Kim, J., and Shaw, E. (2009). "Pedagogical discourse: connecting students to past discussions and peer mentors within an online discussion board," in *Twenty-First IAAI Conference* (Pasadena, MD).

Kim, Y., Lee, H., Shin, J., and Jung, K. (2019). "Improving neural question generation using answer separation," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Honolulu, HI), 6602–6609. doi: 10.1609/aaai.v33i01.33016602

Kloft, M., Stiehler, F., Zheng, Z., and Pinkwart, N. (2014). "Predicting MOOC dropout over weeks using machine learning methods," in *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (Doha), 60–65. doi: 10.3115/v1/W14-4111

Knofczynski, A. (2017). *Why Global Drop-Out Rates Aren't Improving*. The Borgen Project.

Kolb, D. A. (1976). *Learning Style Inventory: Technical Manual*. Boston, MA: McBer.

Koncel-Kedziorski, R., Konstas, I., Zettlemoyer, L., and Hajishirzi, H. (2016). "A theme-rewriting approach for generating algebra word problems," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX), 1617–28. doi: 10.18653/v1/D16-1168

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., et al. (2015). "A machine learning framework to identify students at risk of adverse academic outcomes," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney), 1909–1918. doi: 10.1145/2783258.2788620

Lameras, P., and Arnab, S. (2021). Power to the teachers: an exploratory review on artificial intelligence in education. *nformation* 13, 14. doi: 10.3390/info13010014

Lan, A. S., Vats, D., Waters, A. E., and Baraniuk, R. G. (2015). "Mathematical language processing: automatic grading and feedback for open response mathematical questions," in *Proceedings of the Second (2015) ACM Conference on Learning@ scale* (Vancouver), 167–176. doi: 10.1145/2724660.2724664

Lange, R. C., and Mancoridis, S. (2007). "Using code metric histograms and genetic algorithms to perform author identification for software forensics," in *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation* (London), 2082–2089. doi: 10.1145/1276958.1277364

Larabi-Marie-Sainte, S., Jan, R., Al-Matouq, A., and Alabduhadi, S. (2021). The impact of timetable on student's absences and performance. *PLoS ONE* 16, e0253256. doi: 10.1371/journal.pone.0253256

Li, H., Wang, S., Liang, J., Huang, S., and Xu, B. (2009). "High performance automatic mispronunciation detection method based on neural network and trap features," in *Tenth Annual Conference of the International Speech Communication Association* (Brighton).

Li, Q., and Kim, J. (2021). A deep learning-based course recommender system for sustainable development in education. *Appl. Sci.* 11, 8993. doi: 10.3390/app11198993

Li, W., Li, K., Siniscalchi, S. M., Chen, N. F., and Lee, C.-H. (2016). "Detecting mispronunciations of l2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees," in *Interspeech* (San Francisco, CA), 3127–3131.

Li, X., Chen, M., and Nie, J.-Y. (2020). SEDNN: shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowl. Based Syst.* 210, 106491. doi: 10.1016/j.knosys.2020.106491

Lin, C. F., Yeh, Y.-C., Hung, Y. H., and Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: an application of decision trees. *Comput. Educ.* 68, 199–210. doi: 10.1016/j.compedu.2013.05.009

Lin, L.-H., Chang, T.-H., and Hsu, F.-Y. (2019). "Automated prediction of item difficulty in reading comprehension using long short-term memory," in *2019 International Conference on Asian Language Processing (IALP)* (Shanghai), 132–135. doi: 10.1109/IALP48816.2019.9037716

Lin, M.-H., Chen, H.-G., and Liu, K. S. (2017). A study of the effects of digital learning on learning motivation and learning outcome. *Eur. J. Math. Sci. Technol. Educ.* 13, 3553–3564. doi: 10.12973/eurasia.2017.00744a

Lindberg, D., Popowich, F., Nesbit, J., and Winne, P. (2013). "Generating natural language questions to support learning on-line," in *Proceedings of the 14th European Workshop on Natural Language Generation* (Sofia), 105–114.

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., et al. (2019). EKT: exercise-aware knowledge tracing for student performance prediction. *IEEE Trans. Knowl. Data Eng.* 33, 100–115. doi: 10.1109/TKDE.2019.2924374

Liu, T., Fang, Q., Ding, W., Li, H., Wu, Z., and Liu, Z. (2021). "Mathematical word problem generation from commonsense knowledge graph and equations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4225–4240. doi: 10.18653/v1/2021.emnlp-main.348

Lo, J.-J., and Shu, P.-C. (2005). Identification of learning styles online by observing learners' browsing behaviour through a neural network. *Br. J. Educ. Technol.* 36, 43–55. doi: 10.1111/j.1467-8535.2005.00437.x

Lund, B. D., and Wang, T. (2023). Chatting about chatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*. doi: 10.1108/LHTN-01-2023-0009

Majumder, M., and Saha, S. K. (2015). "A system for generating multiple choice questions: with a novel approach for sentence selection," in *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (Beijing), 64–72. doi: 10.18653/v1/W15-4410

Malik, G., Tayal, D. K., and Vij, S. (2019). "An analysis of the role of artificial intelligence in education and teaching," in *Recent Findings in Intelligent Computing Techniques* eds Sa, P. K., Bakshi, S., Hatzilygeroudis, I. K., and Sahoo, M. N. (Springer), 407–417. doi: 10.1007/978-981-10-8639-7_42

Mallik, S., and Gangopadhyay, A. (2023). Proactive and reactive engagement of artificial intelligence methods for education: a review. *arXiv preprint arXiv:2301.10231*.

Manahi, M. S. (2021). *A deep learning framework for the defection of source code plagiarism using siamese network and embedding models* (Master's thesis). Kulliyyah of Information and Communication Technology, Kuala Lumpur, Malaysia. doi: 10.1007/978-981-16-8515-6_31

Marcinkowski, F., Kieslich, K., Starke, C., and Lünich, M. (2020). "Implications of AI (un-) fairness in higher education admissions: the effects of perceived AI (un-) fairness on exit, voice and organizational reputation," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona), 122–130. doi: 10.1145/3351095.3372867

Mardikyan, S., and Badur, B. (2011). Analyzing teaching performance of instructors using data mining techniques. *Inform. Educ.* 10, 245–257. doi: 10.15388/infedu.2011.17

Mbangula, D. K. (2022). "Adopting of artificial intelligence and development in developing countries: perspective of economic transformation," in *Handbook of Research on Connecting Philosophy, Media, and Development in Developing Countries* eds Okocha, D. O., Onobe, M. J., and Alike, M. N. (IGI Global), 276–288. doi: 10.4018/978-1-6684-4107-7.ch018

Mei, X. Y., Aas, E., and Medgard, M. (2019). Teachers' use of digital learning tool for teaching in higher education: exploring teaching practice and sharing culture. *J. Appl. Res. High. Educ.* 11, 522–537. doi: 10.1108/JARHE-10-2018-0202

Mendis, C., Lahiru, D., Pamudika, N., Madushanka, S., Ranathunga, S., and Dias, G. (2017). "Automatic assessment of student answers for geometric theorem proving questions," in *2017 Moratuwa Engineering Research Conference (MERCon)* (Moratuwa), 413–418. doi: 10.1109/MERCon.2017.7980520

Meuschke, N., Siebeck, N., Schubotz, M., and Gipp, B. (2017). "Analyzing semantic concept patterns to detect academic plagiarism," in *Proceedings of the 6th International Workshop on Mining Scientific Publications* (Toronto), 46–53. doi: 10.1145/3127526.3127535

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., and Del Maestro, R. F. (2020). The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE* 15, e0229596. doi: 10.1371/journal.pone.0229596

Mokbel, B., Gross, S., Paassen, B., Pinkwart, N., and Hammer, B. (2013). "Domain-independent proximity measures in intelligent tutoring systems," in *Educational Data Mining 2013* (Memphis, TN).

Moore, J. S. (1998). An expert system approach to graduate school admission decisions and academic performance prediction. *Omega* 26, 659–670. doi: 10.1016/S0305-0483(98)00008-5

Moscoviz, L., and Evans, D. (2022). *Learning Loss and Student Dropouts During the Covid-19 Pandemic: A Review of the Evidence Two Years After Schools Shut Down*. Center for Global Development.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. (2016). "Generating natural questions about an image," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Berlin), 1802–1813. doi: 10.18653/v1/P16-1170

Mostow, J., and Chen, W. (2009). "Generating instruction automatically for the reading strategy of self-questioning," in *AIED* (Brighton), 465–472.

Mota, J. (2008). "Using learning styles and neural networks as an approach to elearning content and layout adaptation," in *Doctoral Symposium on Informatics Engineering* (Porto).

Mothe, J., and Tanguy, L. (2005). "Linguistic features to predict query difficulty," in *ACM Conference on Research and Development in Information Retrieval, SIGIR, Predicting Query Difficulty-Methods and Applications Workshop* (Salvador), 7–10.

Movellan, J., Eckhardt, M., Virnes, M., and Rodriguez, A. (2009). "Sociable robot improves toddler vocabulary skills," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, CA), 307–308. doi: 10.1145/1514095.1514189

Mridha, K., Jha, S., Shah, B., Damodharan, P., Ghosh, A., and Shaw, R. N. (2022). "Machine learning algorithms for predicting the graduation admission," in *International Conference on Electrical and Electronics Engineering* (Greater Noida: Springer), 618–637. doi: 10.1007/978-981-19-1677-9_55

Muñoz-Najar, A., Gilberto, A., Hasan, A., Cobo, C., Azevedo, J. P., and Akmal, M. (2021). Remote learning during COVID-19: Lessons from today, principles for tomorrow. *World Bank*. doi: 10.1596/36665

Nadeem, F., Nguyen, H., Liu, Y., and Ostendorf, M. (2019). "Automated essay scoring with discourse-aware neural models," in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (Florence), 484–493. doi: 10.18653/v1/W19-4450

Nagatani, K., Zhang, Q., Sato, M., Chen, Y.-Y., Chen, F., and Ohkuma, T. (2019). "Augmenting knowledge tracing by considering forgetting behavior," in *The World Wide Web Conference* (San Francisco, CA), 3101–3107. doi: 10.1145/3308558.3313565

Nakagawa, H., Iwasawa, Y., and Matsuo, Y. (2019). "Graph-based knowledge tracing: modeling student proficiency using graph neural network," in *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (Thessaloniki), 156–163. doi: 10.1145/3350546.3352513

Namatherdhala, B., Mazher, N., and Sriram, G. K. (2022). A comprehensive overview of artificial intelligence tends in education. *Int. Res. J. Modern. Eng. Technol. Sci.* 4.

Nghe, N. T., Janecek, P., and Haddawy, P. (2007). "A comparative analysis of techniques for predicting academic performance," in *2007 37th Annual Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports* (Milwaukee, WI).

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., et al. (2016). "MS MARCO: a human generated machine reading comprehension dataset," in *CoCo@ NIPs* (Barcelona).

Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: a review of trends and approaches for educational technology in a global context. *Int. J. Artif. Intell. Educ.* 25, 177–203. doi: 10.1007/s40593-014-0028-6

Obit, J. H., Landa-Silva, D., Sevaux, M., and Ouelhadj, D. (2011). "Non-linear great deluge with reinforcement learning for university course timetabling," in *Metaheuristics-Intelligent Decision Making, Series Operations Research/Computer Science Interfaces* (Springer), 1–19.

Ohmann, T., and Rahal, I. (2015). Efficient clustering-based source code plagiarism detection using piy. *Knowl. Inform. Syst.* 43, 445–472. doi: 10.1007/s10115-014-0742-2

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Achem, J. A. G., Escamilla, J., and Hosseini, S. (2022). Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Educ. Inform. Technol.* 27, 3891–3933. doi: 10.1007/s10639-021-10751-5

Olney, A. M., D'Mello, S., Person, N., Cade, W., Hays, P., Williams, C., et al. (2012). "Guru: a computer tutor that models expert human tutors," in *International Conference on Intelligent Tutoring Systems* (Chania: Springer), 256–261. doi: 10.1007/978-3-642-30950-2_32

Onan, A. (2020). Mining opinions from instructor evaluation reviews: a deep learning approach. *Comput. Appl. Eng. Educ.* 28, 117–138. doi: 10.1002/cae.22179

Ouyang, F., and Jiao, P. (2021). Artificial intelligence in education: the three paradigms. *Comput. Educ. Artif. Intell.* 2, 100020. doi: 10.1016/j.caeai.2021.100020

Ouyang, F., Zheng, L., and Jiao, P. (2022). Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. *Educ. Inform. Technol.* 1–33. doi: 10.1007/s10639-022-10925-9

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.* 35, 27730–27744.

Özcan, E., Misir, M., Ochoa, G., and Burke, E. K. (2012). "A reinforcement learning: great-deluge hyper-heuristic for examination timetabling," in *Modeling, Analysis, and Applications in Metaheuristic Computing: Advancements and Trends* (IGI Global), 34–55. doi: 10.4018/978-1-4666-0270-0.ch003

Pan, L., Lei, W., Chua, T.-S., and Kan, M.-Y. (2019). Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*. doi: 10.48550/arXiv.1905.08949

Pande, C., Witschel, H. F., Martin, A., and Montecchiari, D. (2021). "Hybrid conversational AI for intelligent tutoring systems," in *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. (Virtual).

Pandey, S., and Karypis, G. (2019). "A self-attentive model for knowledge tracing," in *12th International Conference on Educational Data Mining, EDM 2019* (Montreal: International Educational Data Mining Society), 384–389.

Pantelimon, F.-V., Bologa, R., Toma, A., and Posedaru, B.-S. (2021). The evolution of AI-driven educational systems during the COVID-19 pandemic. *Sustainability* 13, 13501. doi: 10.3390/su132313501

Pask, G. (1976). Styles and strategies of learning. *Br. J. Educ. Psychol.* 46, 128–148. doi: 10.1111/j.2044-8279.1976.tb02305.x

Pavlik, P. I., Jr, and Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cogn. Sci.* 29, 559–586. doi: 10.1207/s15516709cog0000_14

Pavlik, P. I. Jr, Cen, H., and Koedinger, K. R. (2009). "Performance factors analysis–A new alternative to knowledge tracing," in *Proceedings of the 14th International Conference of Artificial Intelligence in Education* (Brighton).

Pedro, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial intelligence in Education: Challenges and Opportunities for Sustainable Development*. UNESCO.

Pereira, J. (2016). "Leveraging chatbots to improve self-guided learning through conversational quizzes," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (Salamanca), 911–918. doi: 10.1145/3012430.3012625

Perikos, I., Grivokostopoulou, F., Kovas, K., and Hatzilygeroudis, I. (2016). Automatic estimation of exercises' difficulty levels in a tutoring system for teaching the conversion of natural language into first-order logic. *Expert Syst*. 33, 569–580. doi: 10.1111/exsy.12182

Persing, I., Davis, A., and Ng, V. (2010). "Modeling organization in student essays," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (Cambridge, MA), 229–239.

Persing, I., and Ng, V. (2013). "Modeling thesis clarity in student essays," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Sofia), 260–269. doi: 10.3115/v1/P14-1144

Persing, I., and Ng, V. (2014). "Modeling prompt adherence in student essays," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Baltimore, MA), 1534–1543.

Persing, I., and Ng, V. (2015). "Modeling argument strength in student essays," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1* (Beijing), 543–552. doi: 10.3115/v1/P15-1053

Pertile, S. d. L., Moreira, V. P., and Rosso, P. (2016). Comparing and combining content-and citation-based approaches for plagiarism detection. *J. Assoc. Inform. Sci. Technol*. 67, 2511–2526. doi: 10.1002/asi.23593

Phandi, P., Chai, K. M. A., and Ng, H. T. (2015). "Flexible domain adaptation for automated essay scoring using correlated linear regression," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon), 431–439. doi: 10.18653/v1/D15-1049

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., et al. (2015a). "Deep knowledge tracing," in *Advances in Neural Information Processing Systems, Vol. 28* (Montreal).

Piech, C., Huang, J., Nguyen, A., Phulsuksombati, M., Sahami, M., and Guibas, L. (2015b). "Learning program embeddings to propagate feedback on student code," in *International conference on machine Learning* (Lille), 1093–1102.

Polozov, O., O'Rourke, E., Smith, A. M., Zettlemoyer, L., Gulwani, S., and Popović, Z. (2015). "Personalized mathematical word problem generation," in *Twenty-Fourth International Joint Conference on Artificial Intelligence* (Buenos Aires).

Pu, Y., Wang, C., and Wu, W. (2020). "A deep reinforcement learning framework for instructional sequencing," in *2020 IEEE International Conference on Big Data (Big Data)* (Virtual), 5201–5208.

Purkayastha, N., and Sinha, M. K. (2021). "Unstoppable study with MOOCs during COVID 19 pandemic: a study," in *Library Philosophy and Practice* (Lincoln, NE: University of Nebraska), 1–12. doi: 10.2139/ssrn.3978886

Qiu, Z., Wu, X., and Fan, W. (2019). "Question difficulty prediction for multiple choice problems in medical exams," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing), 139–148. doi: 10.1145/3357384.3358013

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. doi: 10.18653/v1/D16-1264

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., and Getoor, L. (2013). "Modeling learner engagement in MOOCs using probabilistic soft logic," in *NIPS Workshop on Data Driven Education, Vol. 21* (Lake Tahoe, NV), 62.

Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., and Getoor, L. (2014). "Learning latent engagement patterns of students in online courses," in *Twenty-Eighth AAAI Conference on Artificial Intelligence* (Quebec). doi: 10.1609/aaai.v28i1.8920

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., and Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: a review. *Appl. Sci*. 10, 1042. doi: 10.3390/app10031042

Rawatlal, R. (2017). "Application of machine learning to curriculum design analysis," in *2017 Computing Conference* (London), 1143–1151. doi: 10.1109/SAI.2017.8252234

Reddy, S., Levine, S., and Dragan, A. (2017). "Accelerating human learning with deep reinforcement learning," in *NIPS'17 Workshop: Teaching Machines, Robots, and Humans* (Long Beach, CA), 5–9. doi: 10.15607/RSS.2018.XIV.005

Reich, J., and Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science* 363, 130–131. doi: 10.1126/science.aav7958

Salas-Pilco, S. Z., Xiao, K., and Oshima, J. (2022). Artificial intelligence and new technologies in inclusive education for minority students: a systematic review. *Sustainability* 14, 13572. doi: 10.3390/su142013572

San Pedro, M. O. Z., Baker, R. S., Gowda, S. M., and Heffernan, N. T. (2013). "Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system," in *International Conference on Artificial Intelligence in Education* (Memphis, TN: Springer), 41–50. doi: 10.1007/978-3-642-39112-5_5

Schiff, D. (2021). Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI Soc*. 36, 331–348. doi: 10.1007/s00146-020-01033-8

Shashidhar, V., Pandey, N., and Aggarwal, V. (2015). "Automatic spontaneous speech grading: a novel feature derivation technique using the crowd," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1* (Beijing), 1085–1094. doi: 10.3115/v1/P15-1105

Shen, S., Liu, Q., Chen, E., Huang, Z., Huang, W., Yin, Y., et al. (2021). "Learning process-consistent knowledge tracing," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Singapore), 1452–1460. doi: 10.1145/3447548.3467237

Shermis, M. D., and Burstein, J. C. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Routledge. doi: 10.4324/9781410606860

Shi, Z. R., Wang, C., and Fang, F. (2020). Artificial intelligence for social good: a survey. *arXiv preprint arXiv:2001.01818*. doi: 10.48550/arXiv.2001.01818

Shum, S. J. B., and Luckin, R. (2019). Learning analytics and AI: politics, pedagogy and practices. *Br. J. Educ. Technol*. 50, 2785–2793. doi: 10.1111/bjet.12880

Singh, G., Srikant, S., and Aggarwal, V. (2016). "Question independent grading using machine learning: the case of computer program grading," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 263–272. doi: 10.1145/2939672.2939696

Sinha, T., Li, N., Jermann, P., and Dillenbourg, P. (2014). "Capturing 'attrition intensifying' structural traits from didactic interaction sequences of MOOC learners," in *EMNLP 2014* (Doha), 42. doi: 10.3115/v1/W14-4108

Soleman, S., and Purwarianti, A. (2014). "Experiments on the Indonesian plagiarism detection using latent semantic analysis," in *2014 2nd International Conference on Information and Communication Technology (ICoICT)* (Bandung), 413–418. doi: 10.1109/ICoICT.2014.6914098

Somasundaram, M., Latha, P., and Pandian, S. S. (2020). Curriculum design using artificial intelligence (AI) back propagation method. *Proc. Comput. Sci*. 172, 134–138. doi: 10.1016/j.procs.2020.05.020

Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., and Cheng, M. (2020). "Multi-stage pre-training for automated Chinese essay scoring," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Virtual), 6723–6733. doi: 10.18653/v1/2020.emnlp-main.546

Spector, C. (2022). *New Research Details the Pandemic's Variable Impact on U.S. School Districts*. Stanford News. Available online at: https://news.stanford.edu/2022/10/28/new-research-details-pandemics-impact-u-s-school-districts

Srikant, S., and Aggarwal, V. (2014). "A system to grade computer programming skills using machine learning," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY), 1887–1896. doi: 10.1145/2623330.2623377

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., et al. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assess. Eval. High. Educ*. 41, 721–732. doi: 10.1080/02602938.2016.1164830

Stirling, R., Miller, H., and Martinho-Truswell, E. (2017). Government ai readiness index. *Korea* 4, 7812407479.

Su, J., and Yang, W. (2022). Artificial intelligence in early childhood education: a scoping review. *Comput. Educ*. 2022, 100049. doi: 10.1016/j.caeai.2022.100049

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., et al. (2018). "Exercise-enhanced sequential modeling for student performance prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32* (New Orleans, LA). doi: 10.1609/aaai.v32i1.11864

Sultan, M. A., Bethard, S., and Sumner, T. (2014). "Dls @ cu: Sentence similarity from word alignment," in *SemEval@ COLING* (Dublin), 241–246. doi: 10.3115/v1/S14-2039

Tadesse, S., and Muluye, W. (2020). The impact of COVID-19 pandemic on education system in developing countries: a review. *Open J. Soc. Sci*. 8, 159–170. doi: 10.4236/jss.2020.810011

Taghipour, K., and Ng, H. T. (2016). "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX), 1882–1891. doi: 10.18653/v1/D16-1193

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., and Appleton, J. (2014). "Predicting student risks through longitudinal analysis," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY), 1544–1552. doi: 10.1145/2623330.2623355

Tan, J. S., Goh, S. L., Kendall, G., and Sabar, N. R. (2021). A survey of the state-of-the-art of optimisation methodologies in school timetabling problems. *Expert Syst. Appl.* 165, 113943. doi: 10.1016/j.eswa.2020.113943

Tan, S., Doshi-Velez, F., Quiroz, J., and Glassman, E. (2017). *Clustering Latex Solutions to Machine Learning Assignments for Rapid Assessment*. Available online at: https://finale.seas.harvard.edu/files/finale/files/2017clustering_latex_solutions_to_machine_learning_assignments_for_rapid_assessment.pdf

Taoum, J., Nakhal, B., Bevacqua, E., and Querrec, R. (2016). "A design proposition for interactive virtual tutors in an informed environment," in *International Conference on Intelligent Virtual Agents* (Los Angeles, CA: Springer), 341–350. doi: 10.1007/978-3-319-47665-0_30

Tarcsay, B., Vasić, J., and Perez-Tellez, F. (2022). "Use of machine learning methods in the assessment of programming assignments," in *International Conference on Text, Speech, and Dialogue* (Brno: Springer), 341–350. doi: 10.1007/978-3-031-16270-1_13

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., and Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Proc. Comput. Sci.* 1, 2811–2819. doi: 10.1016/j.procs.2010.08.006

Tong, S., Liu, Q., Huang, W., Hunag, Z., Chen, E., Liu, C., et al. (2020). "Structure-based knowledge tracing: an influence propagation view," in *2020 IEEE International Conference on Data Mining (ICDM)* (Sorrento), 541–550. doi: 10.1109/ICDM50108.2020.00063

Toscher, A., and Jahrer, M. (2010). *Collaborative Filtering Applied to Educational Data Mining*. Washington, DC: KDD cup.

Trakunphutthirak, R., Cheung, Y., and Lee, V. C. (2019). "A study of educational data mining: evidence from a Thai university," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Honolulu, HI), 734–741. doi: 10.1609/aaai.v33i01.3301734

Tschuggnall, M., and Specht, G. (2013). "Detecting plagiarism in text documents through grammar-analysis of authors," in *BTW*, 241–259.

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., and Ragos, O. (2020). Transfer learning from deep neural networks for predicting student performance. *Appl. Sci.* 10:2145. doi: 10.3390/app10062145

Ullah, F., Jabbar, S., and Mostarda, L. (2021). An intelligent decision support system for software plagiarism detection in academia. *Int. J. Intell. Syst.* 36, 2730–2752. doi: 10.1002/int.22399

UNESCO (2021). *When Schools Shut: Gendered Impact of COVID-19 School Closures*. UNESCO Publishing.

UNESCO (2022). *UNESCO's Education Response to COVID-19*.

United Nations (2015). *Goal 4: Quality Education*. United Nations.

Upadhyay, U., De, A., and Gomez Rodriguez, M. (2018). "Deep reinforcement learning of marked temporal point processes," in *Advances in Neural Information Processing Systems, Vol. 31* (Montreal).

Uto, M., and Okano, M. (2020). "Robust neural automated essay scoring using item response theory," in *International Conference on Artificial Intelligence in Education* (Ifrane: Springer), 549–561. doi: 10.1007/978-3-030-52237-7_44

Vani, K., and Gupta, D. (2014). "Using k-means cluster based techniques in external plagiarism detection," in *2014 International Conference on Contemporary Computing and Informatics (IC3I)* (Mysore), 1268–1273. doi: 10.1109/IC3I.2014.7019659

Varga, A., and Ha, L. A. (2010). "WLV: a question generation system for the QGSTEC 2010 task b," in *Proceedings of QG2010: The Third Workshop on Question Generation* (Pittsburgh, PA), 80–83.

Vijayalakshmi, V., Panimalar, K., and Janarthanan, S. (2020). Predicting the performance of instructors using machine learning algorithms. *High Technol. Lett.* 26, 49–54. doi: 10.5373/JARDCS/V12SP4/20201461

Villaverde, J. E., Godoy, D., and Amandi, A. (2006). Learning styles' recognition in e-learning environments with feed-forward neural networks. . *Comput. Assist. Learn.* 22, 197–206. doi: 10.1111/j.1365-2729.2006.00169.x

Vincent-Lancrin, S., and van der Vlies, R. (2020). *Trustworth Artificial Intelligence (AI) in Education: Promises and Challenges*. Organisation for Economic Cooperation and Development.

Vujošević-Janičić, M., Nikolić, M., Tošić, D., and Kuncak, V. (2013). Software verification and graph similarity for automated evaluation of students' assignments. *Inform. Softw. Technol.* 55, 1004–1016. doi: 10.1016/j.infsof.2012.12.005

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., and Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* 104, 106189. doi: 10.1016/j.chb.2019.106189

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *J. Educ. Psychol.* 105, 932. doi: 10.1037/a0031882

Wang, K., and Su, Z. (2015). "Automated geometry theorem proving for human-readable proofs," in *Twenty-Fourth International Joint Conference on Artificial Intelligence* (Buenos Aires).

Wang, L., Sy, A., Liu, L., and Piech, C. (2017). Learning to represent student knowledge on programming exercises using deep learning. *Int. Educ. Data Mining Soc.* doi: 10.1145/3051457.3053985

Wang, T., and Cheng, E. C. (2022). "Towards a tripartite research agenda: a scoping review of artificial intelligence in education research," in *Artificial Intelligence in Education: Emerging Technologies, Models and Applications* (Springer), 3–24. doi: 10.1007/978-981-16-7527-0_1

Wang, T., Su, X., Wang, Y., and Ma, P. (2007). Semantic similarity-based grading of student programs. *Inform. Softw. Technol.* 49, 99–107. doi: 10.1016/j.infsof.2006.03.001

Wang, Z., Lan, A., and Baraniuk, R. (2021). "Math word problem generation with mathematical consistency and problem context constraints," in *2021 Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/2021.emnlp-main.484

Waters, A., and Miikkulainen, R. (2014). Grade: machine learning support for graduate admissions. *AI Mag.* 35, 64. doi: 10.1609/aimag.v35i1.2504

Wen, M., Yang, D., and Rose, C. (2014). "Sentiment analysis in MOOC discussion forums: what does it tell us?," in *Educational Data Mining 2014* (London).

Wester, E. R., Walsh, L. L., Arango-Caro, S., and Callis-Duehl, K. L. (2021). Student engagement declines in stem undergraduates during COVID-19-driven remote learning. *J. Microbiol. Biol. Educ.* 22, ev22i1-2385. doi: 10.1128/jmbe.v22i1.2385

Winthrop, R. (2018). *Leapfrogging Inequality: Remaking Education to Help Young People Thrive*. Brookings Institution Press.

Wolf, S., Aurino, E., Brown, A., Tsinigo, E., and Edro, R. M. (2022). *Remote Data-Collection During COVID 19: Thing of the Past or the Way of the Future*. World Bank Blogs. Available online at: https://blogs.worldbank.org/education/remote-data-collection-during-covid-19-thing-past-or-way-future

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *Int. J. Learn. Technol.* 4, 129–164. doi: 10.1504/IJLT.2009.028804

Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., et al. (2010). "The effect of motivational learning companions on low achieving students and students with disabilities," in *International Conference on Intelligent Tutoring Systems* (Pittsburgh, PA: Springer), 327–337. doi: 10.1007/978-3-642-13388-6_37

Wu, M., Mosse, M., Goodman, N., and Piech, C. (2019). "Zero shot learning for code education: rubric sampling with deep learning inference," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Honolulu, HI), 782–790. doi: 10.1609/aaai.v33i01.3301782

Wu, Q., Zhang, Q., and Huang, X. (2022). Automatic math word problem generation with topic-expression co-attention mechanism and reinforcement learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* 30, 1061–1072. doi: 10.1109/TASLP.2022.3155284

Xie, X., Siau, K., and Nah, F. F.-H. (2020). COVID-19 pandemic-online education in the new normal and the next normal. *J. Inform. Technol. Case Appl. Res.* 22, 175–187. doi: 10.1080/15228053.2020.1824884

Xu, J., Han, Y., Marcu, D., and Van Der Schaar, M. (2017). "Progressive prediction of student performance in college programs," in *Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA).

Xue, K., Yaneva, V., Runyon, C., and Baldwin, P. (2020). "Predicting the difficulty and response time of multiple choice questions using transfer learning," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (Seattle, WA), 193–197. doi: 10.18653/v1/2020.bea-1.20

Yaneva, V., Baldwin, P., and Mee, J. (2020). "Predicting item survival for multiple choice questions in a high-stakes medical exam," in *Proceedings of The 12th Language Resources and Evaluation Conference* (Marseille), 6812–6818.

Yang, D., Sinha, T., Adamson, D., and Rosé, C. P. (2013). "Turn on, tune in, drop out: anticipating student dropouts in massive open online courses," in *Proceedings of the 2013 NIPS Data-Driven Education Workshop, Vol. 11* (Lake Tahoe, NV), 14.

Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., et al. (2020). "GIKT: a graph-based interaction model for knowledge tracing," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Ghent: Springer), 299–315. doi: 10.1007/978-3-030-67658-2_18

Young, N., and Caballero, M. (2019). "Using machine learning to understand physics graduate school admissions," in *Proceedings of the Physics Education Research Conference (PERC)* (Provo, UT), 669–674.

Yudelson, M. V., Koedinger, K. R., and Gordon, G. J. (2013). "Individualized Bayesian knowledge tracing models," in *International Conference on Artificial Intelligence in Education* (Memphis, TN: Springer), 171–180. doi: 10.1007/978-3-642-39112-5_18

Yufeia, L., Salehb, S., Jiahuic, H., and Syed, S. M. (2020). Review of the application of artificial intelligence in education. *Integration* 12:1–5. doi: 10.53333/IJICC2013/12850

Zatarain-Cabada, R., Barrón-Estrada, M. L., Angulo, V. P., García, A. J., and García, C. A. R. (2010). "A learning social network with recognition of learning styles using neural networks," in *Mexican Conference on Pattern Recognition* (Puebla: Springer), 199–209. doi: 10.1007/978-3-642-15992-3_22

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators? *Int. J. Educ. Technol. High. Educ*. 16, 1–27. doi: 10.1186/s41239-019-0171-0

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., et al. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity* 2021, 8812542. doi: 10.1155/2021/8812542

Zhang, D., Wang, L., Zhang, L., Dai, B. T., and Shen, H. T. (2019). The gap of semantic parsing: a survey on automatic math word problem solvers. *IEEE Trans. Pattern Anal. Mach. Intell*. 42, 2287–2305. doi: 10.1109/TPAMI.2019.2914054

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L., et al. (2019). erevise: "Using natural language processing to provide formative feedback on text evidence usage in student writing," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33* (Honolulu, HI), 9619–9625. doi: 10.1609/aaai.v33i01.33019619

Zhang, J., Shi, X., King, I., and Yeung, D.-Y. (2017). "Dynamic key-value memory networks for knowledge tracing," in *Proceedings of the 26th International Conference on World Wide Web* (Perth), 765–774. doi: 10.1145/3038912.3052580

Zhang, L., Zhao, Z., Ma, C., Shan, L., Sun, H., Jiang, L., et al. (2020). End-to-end automatic pronunciation error detection based on improved hybrid CTC/attention architecture. *Sensors* 20, 1809. doi: 10.3390/s20071809

Zhang, M., Baral, S., Heffernan, N., and Lan, A. (2022). "Automatic short math answer grading via in-context meta-learning," in *Proceedings of the International Conference on Educational Data Mining* (Durham).

Zhao, Y., Lackaye, B., Dy, J. G., and Brodley, C. E. (2020). "A quantitative machine learning approach to master students admission for professional institutions," in *International Educational Data Mining Society* (Virtual).

Zhao, Y., Ni, X., Ding, Y., and Ke, Q. (2018). "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 3901–3910. doi: 10.18653/v1/D18-1424

Zhou, Q., and Huang, D. (2019). "Towards generating math word problems from equations and topics," in *Proceedings of the 12th International Conference on Natural Language Generation* (Tokyo), 494–503. doi: 10.18653/v1/W19-8661

Zhu, K., Li, L. D., and Li, M. (2021). A survey of computational intelligence in educational timetabling. *Int. J. Mach. Learn. Comput*. 11, 40–47. doi: 10.18178/ijmlc.2021.11.1.1012

frontiers | Frontiers in Artificial Intelligence

TYPE Systematic Review
PUBLISHED 03 July 2023
DOI 10.3389/frai.2023.1162454

OPEN ACCESS

EDITED BY
Christos Troussas,
University of West Attica, Greece

REVIEWED BY
Zoe Kanetaki,
University of West Attica, Greece
Eliseo Reategui,
Federal University of Rio Grande do Sul, Brazil

*CORRESPONDENCE
Johanna Fleckenstein
✉ fleckenstein@uni-hildesheim.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 20 February 2023
ACCEPTED 30 May 2023
PUBLISHED 03 July 2023

CITATION
Fleckenstein J, Liebenow LW and Meyer J (2023) Automated feedback and writing: a multi-level meta-analysis of effects on students' performance.
Front. Artif. Intell. 6:1162454.
doi: 10.3389/frai.2023.1162454

Automated feedback and writing: a multi-level meta-analysis of effects on students' performance

Johanna Fleckenstein[1,2*†], Lucas W. Liebenow[2†] and Jennifer Meyer[2]

[1]Digital Learning and Instruction, Department of Educational Science, University of Hildesheim, Hildesheim, Germany, [2]Department of Educational Research and Educational Psychology, Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Introduction: Adaptive learning opportunities and individualized, timely feedback are considered to be effective support measures for students' writing in educational contexts. However, the extensive time and expertise required to analyze numerous drafts of student writing pose a barrier to teaching. Automated writing evaluation (AWE) tools can be used for individual feedback based on advances in Artificial Intelligence (AI) technology. A number of primary (quasi-)experimental studies have investigated the effect of AWE feedback on students' writing performance.

Methods: This paper provides a meta-analysis of the effectiveness of AWE feedback tools. The literature search yielded 4,462 entries, of which 20 studies ($k = 84$; $N = 2,828$) met the pre-specified inclusion criteria. A moderator analysis investigated the impact of the characteristics of the learner, the intervention, and the outcome measures.

Results: Overall, results based on a three-level model with random effects show a medium effect ($g = 0.55$) of automated feedback on students' writing performance. However, the significant heterogeneity in the data indicates that the use of automated feedback tools cannot be understood as a single consistent form of intervention. Even though for some of the moderators we found substantial differences in effect sizes, none of the subgroup comparisons were statistically significant.

Discussion: We discuss these findings in light of automated feedback use in educational practice and give recommendations for future research.

KEYWORDS

technology-based learning, automated writing evaluation, writing instruction, feedback, formative assessment, meta-analysis

1. Introduction

Writing is a fundamental, versatile, and complex skill (Graham, 2019; Skar et al., 2022) that is required in a variety of contexts. Shortcomings in writing skills can thus hinder personal, academic, and professional success (Freedman et al., 2016; Graham et al., 2020). A basic aim of educational systems worldwide is to teach students to become competent writers; however, evidence suggests that while some students may achieve this goal, not all do (National Center for Educational Statistics, 2012, 2017; Graham and Rijlaarsdam, 2016). The situation is even further complicated by the fact that there is a large group of students from different language backgrounds who aspire to become competent writers in (English as) a second or foreign language and who are not always able to meet expectations (Fleckenstein et al., 2020a,b; Keller et al., 2020).

Writing skills are influenced by a variety of factors (Graham, 2018). Interindividual differences between writers are especially problematic as students with weak writing skills learn less in all school subjects compared to their more highly skilled classmates

(Graham, 2019). In order to counteract this disadvantage, writing skills need to be promoted more in school. However, educational institutions often lack the time and personnel resources to do this. Graham (2019) reviewed 28 studies on writing instruction at school, identifying major indicators of inadequacy, including the insufficient instructional time devoted to writing (Brindle et al., 2015) and the absence of the use of digital tools for writing (Coker et al., 2016; Strobl et al., 2019; Williams and Beam, 2019).

In addition to high-quality, evidence-based teaching practice, digital technologies can be an asset in the individual promotion of writing skills. Automated writing evaluation (AWE) systems are able to assess students' writing performance, produce individualized feedback, and offer adaptive suggestions for writing improvement. Several individual empirical investigations have already looked into the employment of writing interventions with automated feedback tools, and some have investigated their effect on writing performance—with heterogeneous findings. Relevant moderators of effectiveness, however, have seldom been analyzed. The purpose of this study is to integrate the quantitative empirical literature on the subject of automated feedback interventions with a meta-analytic approach. Beyond the overall effect of automated feedback on student writing, we are particularly interested in moderating effects of learner and treatment characteristics.

2. Theoretical background

2.1. Formative assessment and AWE

Formative assessment serves to provide individualized learning support through a combination of (1) (standardized) learning progress evaluation, (2) individual task-related feedback, and (3) adaptive support for learners (Souvignier and Hasselhorn, 2018; Böhme and Munser-Kiefer, 2020). Implementing formative assessments is a challenge for educational systems, especially when it comes to higher-order competencies that require complex written responses from students. Assessing complex language performance as a necessary basis of individual feedback is a key challenge for teachers (Zhu and Urhahne, 2015; Fleckenstein et al., 2018). Especially judgment biases (e.g., tendencies toward leniency or severity; Jansen et al., 2019, 2021) and the use of simple heuristics in text assessment (e.g., text length; Fleckenstein et al., 2020a,b) can lead to inaccurate judgments of students' performance. Recent technological developments in the field of Artificial Intelligence (AI)—like AWE systems—can assist in the process of formative writing assessment.

The procedure of automatically scoring and evaluating students' written work through machine learning (ML) and natural language processing (NLP) techniques is known as automated writing evaluation (AWE; Bennett and Zhang, 2015). NLP is a subfield of AI that deals with the interaction between computers and humans using natural language. It involves the development of algorithms and systems that can understand, interpret, and generate human language. This includes ML algorithms, which learn from a large dataset of language examples and human ratings. When trained accordingly, AWE systems can evaluate a range of features of written text, including grammar, spelling, clarity, coherence, structure, and content. Based on these text features, they can assign scores to new texts and provide feedback to the writer (AWE feedback; Hegelheimer et al., 2016; Hockly, 2018).

AWE technology is utilized in a variety of educational contexts (Correnti et al., 2022), mainly for summative assessment purposes. Especially high-stakes standardized tests like the Graduate Record Exam (GRE) and the Test of English as a Foreign Language (TOEFL) have been using AWE technology for an automatic evaluation of students' writing (Zhang, 2021). In recent years, many tools have been developed to transfer this technology to low-stakes in-class writing tasks. The two major potentials of AWE with respect to formative assessment in writing are (a) assessment in terms of automatic evaluation of linguistically complex student responses and (b) individualized support through immediate and specific feedback based on students' performance. Various studies have demonstrated the quality of AWE assessment (Shermis, 2014; Perin and Lauterbach, 2018; Rupp et al., 2019; Zawacki-Richter et al., 2019). This review, however, focuses on the second part: Feedback that is based on the automated assessment. In the field of technology-supported writing instruction, this typically means supporting learners by providing adaptive automatic feedback on different textual aspects. While automatic assessment is not the central subject of this meta-analysis, it is the necessary foundation for adaptive feedback and individualized support. Therefore, automated assessment is an important inclusion criterion for the studies considered in this meta-analysis.

2.2. Feedback and AWE

Feedback is generally considered to be one of the most effective factors influencing student learning. This is not only shown by a solid empirical research base ($d = 0.62$; Hattie, 2022) but is also consistent with teachers' professional beliefs (Fleckenstein et al., 2015). For writing feedback in particular, a meta-analysis by Graham et al. (2015) showed effect sizes ranging from $d = 0.38$ to $d = 0.87$, depending on the source of the feedback. Despite these positive findings, process-oriented feedback, in particular, is rarely used by teachers in the classroom as it requires a lot of time and effort (Graham and Hebert, 2011). Feedback has a particularly positive effect on learner performance when it is given in a timely manner when it clarifies the gap between current performance and learning goal, when it reduces cognitive load, and when it is task-related, specific, and detailed (Mory, 2004; Hattie and Timperley, 2007; Shute, 2008; Black and Wiliam, 2009).

In the context of automated text evaluation, the quality of machine judgments is often evaluated on the basis of their agreement with human judgments. In terms of reliability and validity, many studies have come up with satisfactory results in this regard (Shermis, 2014; Rupp et al., 2019; Latifi and Gierl, 2021). Human raters do not necessarily outperform technology in all areas of text evaluation. With respect to segmenting and analyzing texts, experts tend to make coding errors, whereas with respect to recognizing relationships between concepts, human raters have been shown to be superior to technology (Burkhart et al., 2020). Moreover, both human and machine ratings can be affected by judgment bias in that certain text features are disproportionately

included in the judgments (Perelman, 2014; Fleckenstein et al., 2020a,b).

Especially for writing complex and long texts, the evidence of the effectiveness of automated feedback has been described heterogeneously (Stevenson and Phakiti, 2014; McNamara et al., 2015; Strobl et al., 2019). In addition, Graham et al. (2015) noted that few randomized controlled experimental studies had been published. Review articles have either looked at the use of digital technologies in writing instruction in general (Williams and Beam, 2019; Al-Wasy, 2020) or focused on tools and how they work rather than their effectiveness (Allen et al., 2016; Strobl et al., 2019; Deeva et al., 2021).

More recent systematic reviews on the effectiveness of AWE feedback provided an overview of the relevant empirical studies and identified research gaps (Nunes et al., 2021; Fleckenstein et al., 2023). However, they did not quantify the effect of automated feedback on performance and, thus, could not empirically investigate the heterogeneity of effects.

Two very recent meta-analyses have examined the effect of AWE systems on writing performance (Zhai and Ma, 2021; Ngo et al., 2022). Ngo et al. (2022) performed a meta-analysis of AWE systems within the context of second or foreign language education. They found an overall between-group effect size of $g = 0.59$ and investigated several moderating variables, including publication data, population data, and treatment data. Zhai and Ma (2021) also included studies on first language writing in their meta-analysis and found an effect size of $g = 0.86$ for AWE on overall writing quality. However, as outcome measures, the authors included holistic scores only, leaving out individual components of writing performance. The authors found significant moderating effects of educational level, target language learners, and genre of writing.

3. Present study

Our meta-analysis goes beyond the scope of the previous meta-analyses concerning methodological and theoretical considerations. Like Ngo et al. (2022), we used a three-level model with random effects to perform the meta-analysis. However, whereas both previous meta-analyses included post-test data only, we included pre-test performance in the between-group analyses to achieve a more accurate effect size estimation (Morris, 2008). This is especially relevant when drawing on non-randomized primary data (i.e., quasi-experimental designs), for which an equal distribution of pre-test scores across groups cannot be assumed. Furthermore, we used robust variance estimation (RVE) to account for the dependence of effect sizes. Like Zhai and Ma (2021), we included L1 and L2 writers; however, we did not limit the range of outcomes and thus covered holistic and analytic measures of writing performance. We also investigated relevant moderators that have been neglected so far, including the type and level of outcome, the type of control condition, and the time of measurement.

This meta-analysis addresses the two following research questions:

RQ1: What is the overall effect of automated feedback tools on student learning based on an integration of primary studies?

RQ2: To what extent is the effect of automated feedback tools moderated by sample, intervention, and outcome characteristics?

4. Methods

4.1. Inclusion criteria

The analysis of the articles was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) model (Moher et al., 2009). This model provides an evidence-based minimum set of items for reporting reviews and meta-analyses. The selection and coding process for the articles was based on these standards.

In order to be included in the meta-analysis, studies needed to meet all of the pre-specified criteria regarding population, intervention, comparators, outcomes, and study design (PICOS) as specified below:

- Population: Students in primary, secondary and post-secondary education (ISCED level 1-7; UNESCO Institute for Statistics, 2012).
- Intervention: Automated writing evaluation (AWE) providing individualized or adaptive feedback to individual students.
- Comparators: Students receiving no feedback, non-automated feedback (e.g., teacher or peer feedback), or a less extensive form of AWE feedback.
- Outcomes: Writing performance (holistic or analytic) on a revision or transfer task.
- Study design: Experimental or quasi-experimental study designs with at least one treatment condition and one control condition.

Furthermore, studies had to be published in scholarly journals in order to be included. Studies investigating computer-mediated feedback by teachers or peers and studies on constructed responses in the context of short-answer formats were not considered in this meta-analysis.

4.2. Literature search strategy

The literature search was conducted in several literature databases (i.e., Ovid, PsycArticles, PsycInfo, Web of Science Core Collection, and ERIC), using various combinations of keywords: "automated writing evaluation;" "automated essay scoring;" writing + computer-assisted; writing + computer-based; writing + "intelligent tutoring system;" writing + "automated feedback;" writing + "electronic feedback;" writing + digital + feedback; writing + digital + scaffolding.

The literature search yielded in a total of $N = 4{,}462$ reports. After removing duplicates, individual abstracts were screened using the open-source software ASReview (Van de Schoot et al., 2021) for screening prioritization. The tool uses Machine Learning to assist researchers in the process of reviewing large numbers of scientific abstracts. Active learning models iteratively improve their predictions in ordering the abstracts for presentation to the

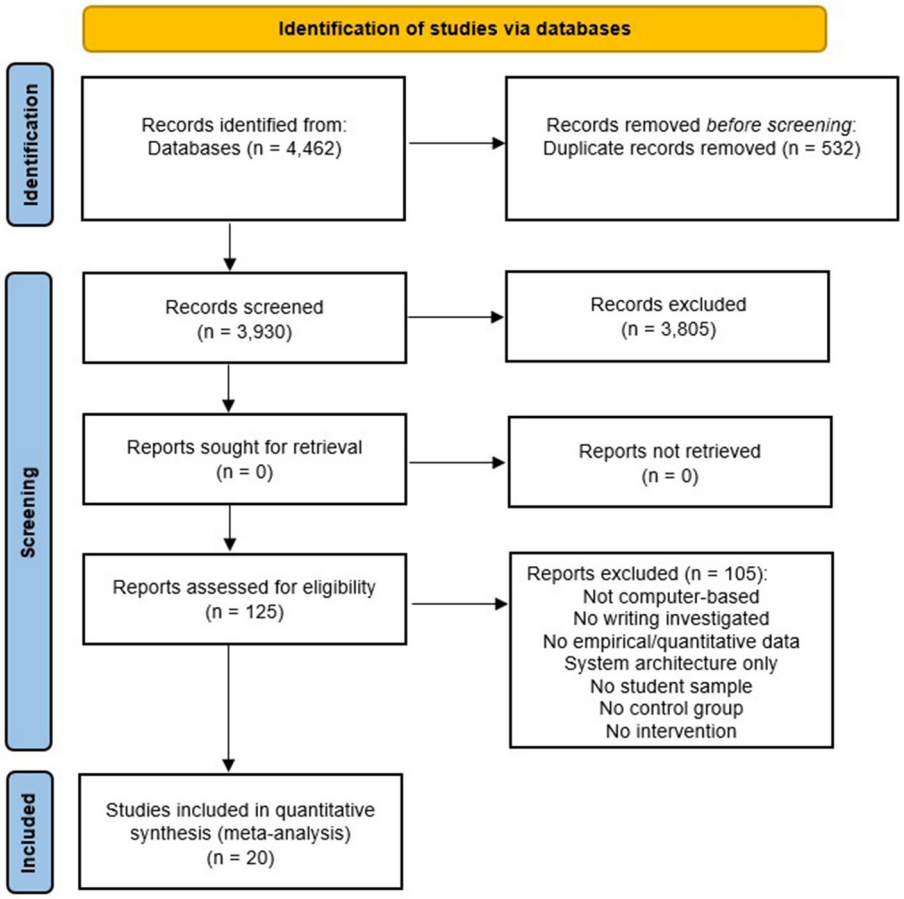


FIGURE 1
Flow-chart of the literature search and screening process (adapted from Moher et al., 2009).

researcher. This procedure has been shown to reduce the number of abstracts to be screened to <40% while retaining a detection rate of 95% of the relevant publications (Ferdinands, 2021). So the goal of ASReview is to help researchers reduce the time and effort required to conduct a literature review, while also improving the quality and comprehensiveness of the review. Based on this, $n = 125$ full texts were screened, identifying $n = 20$ studies that met the inclusion criteria. Figure 1 provides an overview of the literature search and screening process according to the PRISMA guidelines. Following the identification of relevant studies, a coding scheme was developed, and all studies were coded by two independent coders. Any coding that differed was discussed and reviewed by the first co-authors of this paper and corrected if necessary. The variables that were coded and included in the moderator analyses are described in Section 4.5.

4.3. Effect size calculation

The standardized mean differences, also known as Cohen's *d*, between treatment and control conditions were calculated using the R package *esc* (Lüdecke, 2019). For studies that did not report raw statistics (e.g., means and standard deviation), we calculated Cohen's *d* based on other statistical indices (e.g., *F*- or *t*-values).

Morris (2008) recommended an effect size calculation based on the mean pre-post change in the treatment group minus the mean pre-post change in the control group, divided by the pooled pre-test standard deviation. This method was shown to be superior in terms of bias, precision, and robustness to the heterogeneity of variance. Thus, whenever pre-test values were available, they were considered in addition to the post-test values (also see Lipsey and Wilson, 2001; Wilson, 2016; Lüdecke, 2019). In further analyses, we conducted the same model but without considering the corresponding pre-test values to evaluate potential differences in results.

It has been found that Cohen's *d* tends to overestimate the true effect size when the study sample size is small (Grissom and Kim, 2005), which is the case in some of the included primary studies. Therefore, all Cohen's *d* values were converted into Hedges' *g*, which is an unbiased estimator that takes into account the sample sizes (Hedges, 1981):

$$g = 1 - \frac{3}{4\,(n_1 + n_2 - 2) - 1} * d$$

To verbally classify the effect sizes, we used a heuristic derived from the distribution of effects in this research field. This considered the 33rd and 67th percentile of the absolute value of all effects found in this meta-analysis: effects smaller than the 33rd were described as small, effects between the 33rd and 67th

percentiles were described as medium, and effects greater than the 67th percentile were described as large (see Kraft, 2020, for a discussion on how to classify effect sizes; Jansen et al., 2022).

4.4. Meta-analytic integration of effect sizes

We combined the effect sizes of the included studies by applying a three-level model with random effects to take into account that several studies of our meta-analysis reported more than one effect size (Geeraert et al., 2004; Konstantopoulos, 2011; Cheung, 2014; Van den Noortgate et al., 2015; Assink and Wibbelink, 2016). This three-level model considers three levels of variance: variance of the extracted effect sizes at level 1 (sampling variance); variance between effect sizes of a single study at level 2 (within-study variance); and variance between studies at level 3 (between-study variance). Thus, this hierarchical model accounts for the variation of effect sizes between participants (level 1), outcomes (level 2), and studies (level 3).

The multilevel approach is a statistical approach that does not require the correlations between outcomes within primary studies to be known in order to estimate the covariance matrix of the effect sizes. Instead, the second level of the three-level meta-analytic model accounts for sampling covariation (Van den Noortgate et al., 2013). Also, the three-level approach allows for examining differences in outcomes within studies (i.e., within-study heterogeneity) as well as differences between studies (i.e., between-study heterogeneity). If a study reported multiple effect sizes from the same sample that could not be treated as independent from each other, we accounted for this non-independence by using the cluster-robust inference method (also called robust variance estimation; RVE; Sidik and Jonkman, 2006; Hedges et al., 2010; Tipton and Pustejovsky, 2015). This estimation allows for the integration of statistically dependent effect sizes within a meta-analysis without the need for knowledge of the covariance structure among the effect sizes. Furthermore, we conducted moderator analyses to test variables that may reduce within-study or between-study heterogeneity. For these analyses, the three-level random effects model can easily be extended by study and effect size characteristics into a three-level mixed-effects model.

The amount of heterogeneity (i.e., τ^2), was estimated using the restricted maximum-likelihood estimator (Viechtbauer, 2005). In addition to the estimate of τ^2, the Q-test for heterogeneity (Cochran, 1954) and the τ^2 statistic (Higgins and Thompson, 2002) are reported. In case any amount of heterogeneity is detected (i.e., $\tau^2 > 0$, regardless of the results of the Q-test), a prediction interval for the true effect is provided (Riley et al., 2011). The regression test (Sterne and Egger, 2005), using the standard error of the observed outcomes as a predictor, is used to check for funnel plot asymmetry. The analysis was carried out using R (version 4.1.2; R Core Team, 2021) and the *metafor* package (Viechtbauer, 2010) to perform the meta-analyses. In addition, we used the *clubSandwich* package (Pustejovsky, 2022) to perform the cluster-robust inference method.

4.5. Moderation analyses

In combination with the consideration of heterogeneity in our data and calculated effect sizes, we performed several moderator analyses. Moderator variables can be used to provide a more meaningful interpretation of the data and reduce the heterogeneity of the overall effect. First, we identified possible moderator variables from the full texts of the primary studies: sample characteristics (educational level and language status); Intervention characteristics (treatment duration and type of control condition); outcomes characteristics (time of measurement, type of outcome, and outcome level). Second, the $n = 20$ studies included in the meta-analysis were coded by two authors of this study. Third, based on the final codes, the primary studies were divided into subgroups or factors that potentially explain the variance of the observed overall effect. In the following, the coded variables are explained in more detail.

4.5.1. Sample characteristics

4.5.1.1. Educational level

Studies that examined the effect of individual AWE feedback in high school (secondary level) were separated from studies that investigated higher education (tertiary level) students.

4.5.1.2. Language status

As a sample characteristic, we coded language status into L1 for first or majority language contexts and L2 for second or foreign language contexts.

4.5.2. Intervention characteristics

4.5.2.1. Treatment duration

Interventions differed greatly in their duration, ranging from 50 min to one semester. Thus, we categorized intervention duration into short (one or two sessions) and long (more than two sessions).

4.5.2.2. Type of control condition

The studies differed in their design with respect to the control group. In some studies, the control condition received no feedback of any kind on their writing; in other studies, the control condition received a different kind of feedback than the intervention group, such as teacher feedback, peer feedback, or a less extensive form of AWE feedback.

4.5.3. Outcome characteristics

4.5.3.1. Time of measurement

The reported effects were classified as either post-test performance (directly after the intervention) or follow-up performance (time gap between intervention and test.

4.5.3.2. Type of outcome

Most studies on AWE feedback consider either the performance on a text revision or the performance on a different writing task. These outcomes differ in their conceptualization, as a successful revision can be considered performance improvement and a successful transfer to a new task can be considered learning.

TABLE 1 Overall average effect size and heterogeneity test results including pre-test values.

Weighted ES			95% CI		Heterogeneity					
k	*g*	*SE*	Lower	Upper	*Q*	*df*	*p*	I^2_{level3}	I^2_{level2}	I^2_{level1}
84	0.55	0.17	0.19	0.91	285.89	83	<0.001	81.37%	3.85%	14.78%

ES, effect size; CI, confident interval; k, number of effect sizes; g, Hedges' g standardized mean differences; SE, standard error.

TABLE 2 Overall average effect size and heterogeneity test results without pre-test values.

Weighted ES			95% CI		Heterogeneity					
k	*g*	*SE*	Lower	Upper	*Q*	*df*	*p*	I^2_{level3}	I^2_{level2}	I^2_{level1}
84	0.77	0.20	0.35	1.18	985.01	83	<0.000	85,55%	9.71%	4.74%

ES, effect size; CI, confident interval; k, number of effect sizes; g, Hedges' g standardized mean differences; SE, standard error.

4.5.3.3. Outcome level

Furthermore, outcomes were categorized according to the level of detail. Outcomes were considered holistic when the effect referred to a total score or grade for the whole text. Analytic outcomes were further differentiated for effects concerning language aspects (e.g., grammar and mechanics) or content aspects (e.g., unity and number of subthemes) of the text.

5. Results

5.1. Overall effect of AWE feedback

A total of $k = 84$ effect sizes involving $N = 2{,}828$ learners from 20 studies were included in the analysis. The observed effects ranged from −1.73 to 2.99, with the majority of estimates being positive (70.24%). The estimated average effect size based on the three-level model with random effects was $g = 0.55$ ($SE = 0.17$) and differed significantly from zero ($z = 3.18$, $p < 0.001$). The comparison of the three-level model with the conventionally two-level model showed a significantly better fit for the three-level model, based on the likelihood ratio test ($X^2 = 150.36$; $p < 0.001$). Therefore, the application of the three-level model would better explain the between-group comparison data.

According to the *Q*-test, the effect showed significant heterogeneity (see Table 1). The estimated variance values were τ^2 for level 3 = 0.51 and τ^2 for level 2 = 0.02. A 95% prediction interval for the estimated effect is given by −1.02 to 2.12. Hence, although the average effect is estimated to be positive, in some studies, the true effect may, in fact, be negative.

As a further analysis, we examined whether the overall effect sizes differ when we ignore pre-test values from primary studies that provided them and calculated the effect sizes based only on the post-test values from the studies in concern (see Table 2). We observed an estimated average effect of $g = 0.77$ ($SE = 0.20$). The observed effects ranged from −1.14 to 3.61, with the majority of estimates being positive (70.24%). Therefore, the average effect differed significantly from zero ($z = 3.87$, $p < 0.001$). However, a Wald test showed that both effect sizes did not significantly differ from each other ($Q = 0.67$, $p = 0.414$).

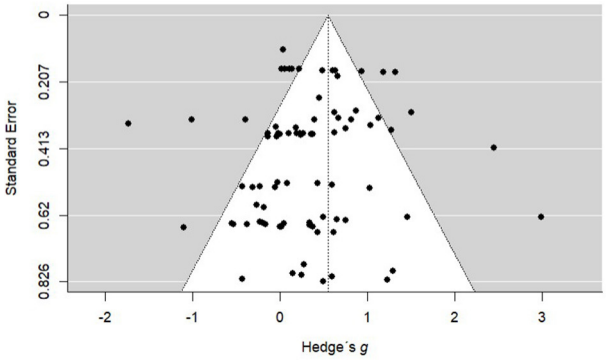


FIGURE 2
Funnel plot of all study effects.

5.2. Publication bias

To examine a publication bias, we used a funnel plot to see whether there is a symmetry of effect sizes, as they should be evenly distributed on both sides of the centered line, which represents the overall average effect sizes across all unique samples (Figure 2). In addition, we ran an Egger's test to evaluate the statistical significance of the asymmetry of the funnel plot by using the squared standard errors of the effect size estimates as a predictor in the meta-regression (Sterne and Egger, 2005). The results of the test confirmed that our funnel plot asymmetry is not different from zero ($b = -0.76$, SE = 0.69, $z = -1.09$, $p = 0.27$, 95% *CI* [−2.12 - 0.60]), indicating that there are no conspicuous data characteristics, producing an asymmetric inverted funnel plot. Therefore, we can assume the absence of a significant publication bias.

5.3. Moderation analysis

To test our hypotheses, we computed a random effects model with subgroup and regression effects of our coded moderator variables (see Table 3). For verbal classification, we used the 33rd percentile ($g = 0.23$) and the 67th percentile ($g = 0.60$) of the absolute values of the effects. Thus, effects below $g = 0.23$ were classified as "small,"

TABLE 3 Moderation effects with *z*-tests against zero.

Moderator	*k*	*g*	*SE*	*z*	*df*	*p*	95% CI
Sample characteristics							
Educational level							
Secondary	23	0.50	0.16	3.13	4.95	0.03	[0.089; 0.921]
Tertiary	61	0.58	0.25	2.33	12.62	0.04	[0.039; 1.114]
Language status							
L1	58	0.40	0.12	3.28	8.86	0.01	[0.124; 0.676]
L2	26	0.72	0.35	2.08	8.71	0.07	[−0.066; 1.506]
Intervention characteristics							
Treatment duration							
Long	47	0.66	0.22	3.04	14.54	0.01	[0.196; 1.117]
Short	37	0.18	0.11	1.65	3.00	0.20	[−0.163; 0.514]
Type of control condition							
No feedback	68	0.59	0.15	3.84	14.70	0.00	[0.261; 0.912]
Other feedback	16	0.40	0.71	0.57	2.85	0.61	[−1.925; 2.73]
Outcome characteristics							
Time of measurement							
Post	66	0.57	0.17	3.44	18.43	0.00	[0.224; 0.926]
Follow-up	18	0.27	0.27	0.97	5.64	0.37	[−0.415; 0.947]
Type of outcome							
Performance	40	0.27	0.18	1.49	4.54	0.20	[−0.211; 0.755]
Learning	44	0.65	0.21	3.12	13.75	0.01	[0.203; 1.101]
Outcome level							
Holistic	29	0.50	0.22	2.30	16.91	0.04	[0.04; 0.965]
Content	25	0.57	0.18	3.20	13.84	0.01	[0.188; 0.952]
Language	30	0.61	0.17	3.61	12.00	0.00	[0.241; 0.975]

effects between $g = 0.23$ and $g = 0.60$ were classified as "medium," and effects above $g = 0.60$ were classified as "large." The moderators were grouped into three categories: *sample characteristics*, *intervention characteristics*, and *outcome characteristics*.

5.3.1. Sample characteristics

5.3.1.1. Educational level

We found medium effects for both secondary level (0.50) and tertiary level (0.58) that were both significantly different from zero. The difference between the two effects was not statistically significant ($Q = 0.03$, $p = 0.854$).

5.3.1.2. Language status

For samples with the target language as L1, we found a medium effect (0.40); for those with an L2 background, the effect can be categorized as large (0.72). Both effects significantly differed from zero. The effects did not significantly differ from each other ($Q = 0.82$, $p = 0.365$).

5.3.2. Intervention characteristics

5.3.2.1. Treatment duration

Long interventions of more than two sessions showed a large significant effect (0.66), whereas short interventions of one or two sessions showed a small non-significant effect (0.18). However, the difference between the two effects was not statistically significant ($Q = 1.32$, $p = 0.250$).

5.3.2.2. Type of control condition

For those Intervention groups that were compared to a control condition without any kind of feedback, the effect was significant and of medium size (0.59). When compared to a group with a different kind of feedback, the medium effect (0.40) was not significantly different from zero. However, the difference between effects was not statistically significant ($Q = 0.16$, $p = 0.684$).

5.3.3. Outcome characteristics

5.3.3.1. Time of measurement

For both post-test performance (0.57) and follow-up performance (0.27), we found effects that fall in the medium

category. However, the follow-up effect did not significantly differ from zero, whereas the effect on post-test performance did. The difference between the effects was marginally significant on the 10%-level ($Q = 2.71$, $p = 0.099$).

5.3.3.2. Type of outcome

The medium effect (0.27) for revision tasks as the outcome (performance) was not significantly different from zero. For transfer tasks (learning), the effect was large and significant (0.65). Again, the difference between effects was not statistically significant ($Q = 1.75$, $p = 0.186$).

5.3.3.3. Outcome level

The effects of the three outcomes were all of medium-large size (holistic: 0.50; content: 0.57; language: 0.61), and they were all significantly different from zero. The three effects did not significantly differ from each other ($Q = 0.51$, $p = 0.773$).

6. Discussion

In the following, we discuss the central findings of this meta-analysis. Before we provide insight into automated feedback use in educational practice and give recommendations for future research, we briefly summarize our findings regarding the overall effect of AWE feedback and the moderator analyses.

6.1. Summary

This meta-analysis examined the overall effect of AWE feedback on writing performance by collecting 84 effect sizes from 20 primary studies with a total of 2,828 participants. A medium effect size of $g = 0.55$ was obtained using a three-level random-effects model. The findings support the use of AWE feedback to facilitate students' writing development.

The effect size is in line with prior meta-analytic research by Ngo et al. (2022), who found an overall between-group effect of $g = 0.59$. However, it is considerably smaller than the effect of $g = 0.86$ found by Zhai and Ma (2022). This variance in effect sizes may be due to the fact that the latter meta-analysis did not use a three-level model for their data analysis. Thus, they did not account for the dependence of effects reported within one study. They also did not include pre-test performance in their model; however, neither did Ngo et al. (2022).

Our robustness check showed that neglecting the pre-test performance in this research area could lead to an overestimation of the overall effect size ($g = 0.77$). However, this effect—although verbally classified as a large effect—did not significantly differ from the medium effect found in the original analysis.

Since the data showed significant heterogeneity, we investigated the impact of several potential moderators, including characteristics of the sample, the intervention, and the outcome. Even though for some of the moderators, we found substantial differences in effect sizes, none of the subgroup comparisons were statistically significant. This should be kept in mind when verbally classifying the effect sizes. In the following, we interpret our findings in light of previous research, especially the two recent meta-analyses by Ngo et al. (2022) and Zhai and Ma (2022).

Sample characteristics included the educational level and the language context. We differentiated for secondary and tertiary level, finding similar effects of medium size for both. This is contrary to the findings by Ngo et al. (2022) and Zhai and Ma (2022), who both found larger effects for post-secondary learners compared to learners at secondary level. However, both previous meta-analyses only included a very limited number of primary studies drawing on secondary-level samples ($k = 3$ resp. $k = 6$). Thus, it can be assumed that AWE feedback is similarly effective in both contexts. In terms of language context, the effect was large for L2 and medium for L1 contexts. Zhai and Ma (2022) reported a similar finding when comparing learners of English as a second or foreign language with native English speakers.

We found a large effect for long-term AWE feedback treatments but only a small effect for short interventions. This is in line with Ngo et al. (2022), who even found a small negative effect for short durations ($\leq$2 weeks). The difference between medium and long intervention durations in Zhai and Ma (2022), however, was also not statistically significant. Zhai and Ma (2022) did also not find a significant effect for feedback combination (AWE only vs. AWE + teacher vs. AWE + peer). We took a slightly different approach and investigated different control conditions, some of which did not receive any feedback treatment and some of which received a different feedback treatment (e.g., teacher or peer feedback). Contrary to expectations, the medium-size effects did not differ significantly for the two types of control conditions.

Even though many studies in this field report post-test as well as follow-up outcomes, neither of the two prior meta-analyses investigated this as a moderator. We found the overall effect on post-test performance to be of medium size and not significantly different from zero; the effect on follow-up performance was small and did not significantly differ from zero. Again, in direct comparison, the difference between effects did not reach statistical significance. Neither of the previous meta-analyses looked into the type of outcome (i.e., performance vs. learning), even though this is a striking difference between studies that could explain the heterogeneity. To our surprise, the effect for revision tasks (performance improvement) was small, whereas the effect for transfer tasks (learning) was large. Only the latter differed significantly from zero. This indicates that AWE feedback does have an impact on learning to write rather than on situational performance enhancement. Unfortunately, the number of studies available does not suffice to investigate interactions of type of outcome with other moderator variables. Zhai and Ma (2022) only investigate holistic text quality as an outcome. Ngo et al. (2022) investigated outcome measure as a moderator with seven categories, finding effect sizes that ranged from $g = 0.27$ (Grammar and Mechanics) to $g = 0.83$ (Vocabulary). However, the effect sizes did not significantly differ from each other, probably due to small subgroup sizes. In our analysis of holistic and analytic (content, language) outcomes, we found very similar effects of medium size. More research on outcome measures as moderators of AWE feedback effectiveness is needed to investigate differential effects more closely.

6.1.1. Limitations and directions for future research

Even though effects differed in size for some of the moderators, these differences were not statistically significant. Thus, the detected heterogeneity may be explained by variables other than the ones that we attended to in our moderator analyses. Thus, in future research, additional moderators need to be investigated. In other learning contexts, the type of feedback has been shown to moderate effectiveness (Van der Kleij et al., 2015; Wisniewski et al., 2020; Mertens et al., 2022). In the context of AWE feedback, we need more primary studies that compare different types of feedback or at least provide sufficient information on the details of their feedback intervention. Moreover, the design and presentation of automated feedback have rarely been investigated (for an exception, see Burkhart et al., 2020).

The potential to identify publication bias in a certain area of research is one of the strengths of meta-analytic research. We assessed publication bias by testing the asymmetry of the funnel plot, finding no indicator for bias. However, a more thorough analysis of publication bias is needed. In order to find out whether non-significant or small effects of AWE feedback tend to remain unpublished, the respective meta-analysis should include unpublished or non-peer-reviewed primary studies.

The variance in estimated effect sizes across AWE feedback meta-analyses calls for a second-order meta-analysis. The purpose of a second-order meta-analysis is to estimate the proportion of the variance in meta-analytic effect sizes across multiple first-order meta-analyses attributable to second-order sampling error and to use this information to improve the accuracy of estimation for each first-order meta-analytic estimate (Schmidt and Oh, 2013). Thus, a second-order meta-analysis would inform AWE feedback research and provide a more comprehensive understanding of factors influencing AWE feedback effectiveness.

6.2. Practical implications

This meta-analysis showed that AWE feedback has a medium positive effect on students' writing performance in educational contexts. However, the heterogeneity in the data suggests that automated feedback should not be seen as a one-size-fits-all solution, and its impact may vary based on factors such as context and learner characteristics, the feedback intervention itself, and outcome measures.

For teachers and school administrators, this implies that AWE feedback can be a useful tool to support students' writing in educational contexts, but its use should be carefully considered and integrated into a comprehensive approach to writing instruction. The use of automated feedback should be combined with other forms of support, such as teacher feedback and individualized learning opportunities, to ensure its effectiveness.

Furthermore, the heterogeneity in the results suggests that automated feedback may not have the same impact on all students. Teachers and administrators should consider the individual needs and characteristics of their students when deciding whether and how to implement automated feedback. Further research is needed to determine the most effective use of automated feedback in different educational contexts and with different populations. Teachers and administrators should keep up to date with developments in the field and use evidence-based practices to inform their decisions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JF, LL, and JM contributed to conception and design of the study. JF and LL conducted the literature search, screening procedure, and coding. LL performed the statistical analysis and wrote sections of the manuscript. JF wrote the first draft of the manuscript. All authors have read and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allen, L. K., Jacovina, M. E., and McNamara, D. S. (2016). "Computer-based writing instruction," in *Handbook of Writing Research*, eds C. A. MacArthur, S. Graham, and J. Fitzgerald (The Guilford Press), pp. 316–329.

Al-Wasy, B. Q. (2020). The effectiveness of integrating technology in EFL/ESL writing: A meta-analysis. *Interact. Technol. Smart Educ.* 2020, 33. doi: 10.1108/ITSE-03-2020-0033

Assink, M., and Wibbelink, C. J. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *Quantit. Methods Psychol.* 12, 154–174. doi: 10.20982/tqmp.12.3.p154

Bennett, R. E., and Zhang, M. (2015). "Validity and automated scoring," in *Technology and Testing*, ed F. Drasgow (London: Routledge), 142–173.

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educat. Assess. Eval. Accountabil.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Böhme, R., and Munser-Kiefer, M. (2020). Lernunterstützung mit digitalen Unterrichtsmaterialien: Interdisziplinäre Erkenntnisse und Entwicklungsperspektiven. *MedienPädagogik* 17, 427–454. doi: 10.21240/mpaed/jb17/2020.05.17.X

Brindle, M., Graham, S., Harris, K. R., and Hebert, M. (2015). Third and fourth grade teacher's classroom practices in writing: A national survey. *Read. Writ.* 29, 929–954. doi: 10.1007/s11145-015-9604-x

Burkhart, C., Lachner, A., and Nückles, M. (2020). Assisting students' writing with computer-based concept map feedback: A validation study of the CohViz feedback system. *PLoS ONE*. 15, e0235209. doi: 10.1371/journal.pone.0235209

*Cheng, G. (2017). The impact of online automated feedback on students' reflective journal writing in an EFL course. *Internet High. Educ.* 34, 18–27. doi: 10.1016/j.iheduc.2017.04.002

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychol. Methods* 19, 211. doi: 10.1037/a0032968

*Chew, C. S., Idris, N., Loh, E. F., Wu, W. C. V., Chua, Y. P., and Bimba, A. T. (2019). The effects of a1 theory-based summary writing tool on students' summary writing. *J. Comput. Assist. Learn.* 35, 435–449. doi: 10.1111/jcal.12349

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10, 101–129. doi: 10.2307/3001666

Coker, D. L., Farley-Ripple, E., Jackson, A. F., Wen, H., MacArthur, C. A., and Jennings, A. S. (2016). Writing instruction in first grade: An observational study. *Read. Writ.* 29, 793–832. doi: 10.1007/s11145-015-9596-6

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., and Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Comput Educ. Open.* doi: 10.1016/j.caeo.2022.100084

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., and De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Comput Educ.* 162. doi: 10.1016/j.compedu.2020.104094

Ferdinands, G. (2021). AI-assisted systematic reviewing: Selecting studies to compare bayesian versus frequentist SEM for small sample sizes. *Multivar. Behav. Res.* 56, 153–154. doi: 10.1080/00273171.2020.1853501

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R., and Köller, O. (2020a). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assess. Writ.* 43, 100420. doi: 10.1016/j.asw.2019.100420

Fleckenstein, J., Leucht, M., and Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Lang. Assess. Quart.* 15, 90–101. doi: 10.1080/15434303.2017.1421956

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., and Köller, O. (2020b). Is a long essay always a good essay? The effect of text length on writing assessment. *Front. Psychol.* 11, 562462. doi: 10.3389/fpsyg.2020.562462

Fleckenstein, J., Reble, R., Meyer, J., Jansen, T., Liebenow, L. W., Möller, J., et al. (2023). "Digitale Schreibförderung im Bildungskontext: Ein systematisches Review," in *Bildung für eine digitale Zukunft, Vol. 15*, eds K. Scheiter, and I. Gogolin (Wiesbaden: Springer VS), 3–25. doi: 10.1007/978-3-658-37895-0_1

Fleckenstein, J., Zimmermann, F., Köller, O., and Möller, J. (2015). What works in school? Expert and novice teachers' beliefs about school effectiveness. *Frontl. Learn. Res.* 3, 27–46. doi: 10.14786/flr.v3i2.162

Freedman, S. W., Hull, G. A., Higgs, J. M., and Booten, K. P. (2016). Teaching writing in a digital and global age: Toward access, learning, and development for all. *Am. Educ. Res. Assoc.* 6, 23. doi: 10.3102/978-0-935302-48-6_23

*Gao, J., and Ma, S. (2019). The effect of two forms of computer-automated metalinguistic corrective feedback. *Lang. Learn. Technol.* 23, 65–83.

Geeraert, L., Van den Noortgate, W., Grietens, H., and Onghena, P. (2004). The effects of early prevention programs for families with young children at risk for physical child abuse and neglect: A meta-analysis. *Child Maltreat.* 9, 277–291. doi: 10.1177/1077559504264265

Graham, S. (2018). A revised writer (s)-within-community model of writing. *Educ. Psycholog.* 53, 258–279. doi: 10.1080/00461520.2018.1481406

Graham, S. (2019). Changing how writing is taught. *Rev. Res. Educ.* 43, 277–303. doi: 10.3102/0091732X18821125

Graham, S., and Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harv. Educ. Rev.* 81, 710–744. doi: 10.17763/haer.81.4.t2k0m13756113566

Graham, S., Hebert, M., and Harris, K. R. (2015). Formative assessment and writing. *Element. School J.* 115, 523–547. doi: 10.1086/681947

Graham, S., Kiuhara, S. A., and MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Rev. Educ. Res.* 90, 179–226. doi: 10.3102/0034654320914744

Graham, S., and Rijlaarsdam, G. (2016). Writing education around the globe: Introduction and call for a new global analysis. *Read. Writ.* 29, 781–792. doi: 10.1007/s11145-016-9640-1

Grissom, R. J., and Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

*Hassanzadeh, M., and Fotoohnejad, S. (2021). Implementing an automated feedback program for a foreign language writing course: A learner-centric study: Implementing an AWE tool in a L2 class. *J. Comput. Assist. Learn.* 37, 1494–1507. doi: 10.1111/jcal.12587

Hattie, J. (2009). The black box of tertiary assessment: An impending revolution. *Tertiary Assess. High. Educ. Stud. Outcomes* 259, 275.

Hattie, J. (2022). *Visible Learning Meta^x: Feedback*. Available online at: http://www.visiblelearningmetax.com/Influences (accessed January 20, 2023).

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *J. Educ. Stat.* 6, 107–128. doi: 10.3102/10769986006002107

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Methods* 1, 39–65. doi: 10.1002/jrsm.5

Hegelheimer, V., Dursun, A., and Li, Z. (2016). Automated writing evaluation in language teaching: Theory, development, and application. *CALICO J.* 33, I–V. doi: 10.1558/cj.v33i1.29251

Higgins, J. P., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21, 1539–1558. doi: 10.1002/sim.1186

Hockly, N. (2018). Automated writing evaluation. *ELT J.* 73, 82–88. doi: 10.1093/elt/ccy044

Jansen, T., Meyer, J., Wigfield, A., and Möller, J. (2022). Which student and instructional variables are most strongly related to academic motivation in K-12 education? A systematic review of meta-analyses. *Psychol. Bullet.* 148, 1–26. doi: 10.1037/bul0000354

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., and Möller, J. (2021). Judgment accuracy in experienced vs. student teachers: Assessing essays in english as a foreign language. *Teach. Teacher Educ.* 97, 103216. doi: 10.1016/j.tate.2020.103216

Jansen, T., Vögelin, C., Machts, N., Keller, S., and Möller, J. (2019). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen. *Psychologie in Erziehung Und Unterricht* 2019, art21d. doi: 10.2378/peu2019.art21d

Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., and Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *J. Second Lang. Writ.* 48, 100700. doi: 10.1016/j.jslw.2019.100700

*Kellogg, R. T., Whiteford, A. P., and Quinlan, T. (2010). Does automated feedback help students learn to write? *J. Educ. Comput. Res.* 42, 173–196. doi: 10.2190/EC.42.2.c

Klein, P., and Boscolo, P. (2016). Trends in research on writing as a learning activity. *J. Writ. Res.* 7, 311–350. doi: 10.17239/jowr-2016.07.03.01

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Res. Synth. Methods* 2, 61–76. doi: 10.1002/jrsm.35

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educ. Research.* 49, 241–253. doi: 10.3102/0013189X20912798

*Lachner, A., Burkhart, C., and Nückles, M. (2017). Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. *J. Exp. Psychol.* 23, 29. doi: 10.1037/xap0000111

Latifi, S., and Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Lang. Test.* 38, 62–85. doi: 10.1177/0265532220929918

Light, R. J. (2001). *Making the Most of College*. Cambridge, MA: Harvard University Press.

*Lin, M. P. C., and Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. *J. Educ. Technol. Soc.* 23, 78–92.

*Link, S., Mehrzad, M., and Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Comput. Assist. Lang. Learn.* 35, 605–634. doi: 10.1080/09588221.2020.1743323

Lipsey, M. W., and Wilson, D. B. (2001). *Practical Meta-analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

*Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in china CET. *Big Data* 7, 121–129. doi: 10.1089/big.2018.0151

Lüdecke, D. (2019). *ESC: Effect Size Computation for Meta-analysis (Version 0.5.1)*. doi: 10.5281/zenodo.1249218

*McCarthy, K. S., Roscoe, R. D., Likens, A. D., and McNamara, D. S. (2019). "Checking it twice: Does adding spelling and grammar checkers improve essay quality in an automated writing tutor?" in *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29. 2019, Proceedings, Part i 20*, Chicago, IL, 270–282. doi: 10.1007/978-3-030-23204-7_23

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assess. Writ.* 23, 35–59. doi: 10.1016/j.asw.2014.09.002

Mertens, U., Finn, B., and Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *J. Educ. Psychol.* 114, edu0000764. doi: 10.1037/edu0000764

*Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann. Internal Med.* 151, 264–269. doi: 10.7326/0003-4819-151-4-200908180-00135

*Mørch, A. I., Engeness, I., Cheng, V. C., Cheung, W. K., and Wong, K. C. (2017). EssayCritic: Writing to learn with a knowledge-based design critiquing system. *J. Educ. Technol. Soc.* 20, 213–223.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organ. Res. Methods* 11, 364–386. doi: 10.1177/1094428106291059

Mory, E. (2004). "Feedback Research Revisited," in *Handbook of Research on Educational, Communications and Technology*, eds D. Jonasson and M. Driscoll (New York, NY: Routledge), pp. 745–783.

National Center for Educational Statistics (2012). *The Nation's Report Card: Writing 2011*. Available online at: https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf (accessed January 20, 2023).

National Center for Educational Statistics (2017). *Technical Summary of Preliminary Analyses of NAEP 2017 Writing Assessments*. Available online at: https://nces.ed.gov/nationsreportcard/subject/writing/pdf/2017_writing_technical_summary.pdf (accessed January 20, 2023).

Ngo, T. T. N., Chen, H. H. J., and Lai, K. K. W. (2022). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interact. Learn. Environ.* 2022, 1–18. doi: 10.1080/10494820.2022.2096642

Nunes, A., Cordeiro, C., Limpo, T., and Castro, S. L. (2021). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *J. Comput. Assist. Learn.* 38, 599–620. doi: 10.1111/jcal.12635

*Palermo, C., and Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contempor. Educ. Psychol.* 54, 255–270. doi: 10.1016/j.cedpsych.2018.07.002

Perelman, L. (2014). When "the state of the art" is counting words. *Assess. Writ.* 21, 104–111. doi: 10.1016/j.asw.2014.05.001

Perin, D., and Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *Int. J. Artif. Intellig. Educ.* 28, 56–78. doi: 10.1007/s40593-016-0122-z

Pustejovsky, J. (2022). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators With Small-Sample Corrections*. Available online at: https://CRAN.R-project.org/package=clubSandwich (accessed January 20, 2023).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed January 20, 2023).

*Reynolds, B. L., Kao, C.-W., and Huang, Y. (2021). Investigating the effects of perceived feedback source on second language writing performance: A quasi-experimental study. *Asia-Pacific Educ. Research.* 30, 585–595. doi: 10.1007/s40299-021-00597-3

*Riedel, E., Dexter, S. L., Scharber, C., and Doering, A. (2006). Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *J. Educ. Comput. Res.* 35, 267–287. doi: 10.2190/U552-M54Q-5771-M677

Riley, R. D., Higgins, J. P., and Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *Br. Med. J.* 342, d549. doi: 10.1136/bmj.d549

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., and Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Res. Rep. Ser.* 2019, 1–23. doi: 10.1002/ets2.12249

Schmidt, F. L., and Oh, I. S. (2013). Methods for second order meta-analysis and illustrative applications. *Org. Behav. Hum. Decision Process.* 121, 204–218. doi: 10.1016/j.obhdp.2013.03.002

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assess. Writ.* 20, 53–76. doi: 10.1016/j.asw.2013.04.001

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Sidik, K., and Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Comput. Stat. Data Anal.* 50, 3681–3701. doi: 10.1016/j.csda.2005.07.019

Skar, G. B., Graham, S., and Rijlaarsdam, G. (2022). Formative writing assessment for change—introduction to the special issue. *Assess. Educ. Principl. Pol. Practice* 29, 121–126. doi: 10.1080/0969594X.2022.2089488

Souvignier, E., and Hasselhorn, M. (2018). Formatives assessment. *Zeitschrift für Erziehungswissenschaft.* 21, 693–696. doi: 10.1007/s11618-018-0839-6

Sterne, J. A., and Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. *Publicat. Bias Meta-Analysis* 6, 99–110. doi: 10.1002/0470870168.ch6

Stevenson, M., and Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assess. Writ.* 19, 51–65. doi: 10.1016/j.asw.2013.11.007

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., et al. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Comput. Educ.* 131, 33–48. doi: 10.1016/j.compedu.2018.12.005

*Tang, J., and Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from china. *JALT CALL J.* 13, 117–146. doi: 10.29140/jaltcall.v13n2.215

Tipton, E., and Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *J. Educ. Behav. Stat.* 40, 604–634. doi: 10.3102/1076998615606099

UNESCO Institute for Statistics (2012). *International Standard Classification of Education: ISCED 2011*. Available online at: https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf (accessed January 20, 2023).

Van de Schoot, R., de Bruin, J., de Schram, R., Zahedi, P., de Boer, J., de Weijdema, F., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Machine Intellig.* 3, 125–133. doi: 10.1038/s42256-020-00287-7

Van den Noortgate, W., López-López, J. A., Marin-Martinez, F., and Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behav. Res. Methods* 45, 576–594. doi: 10.3758/s13428-012-0261-6

Van den Noortgate, W., López-López, J. A., Marin-Martinez, F., and Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behav. Res. Methods* 47, 1274–1294. doi: 10.3758/s13428-014-0527-2

Van der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 30, 261–293. doi: 10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03

*Wade-Stein, D., and Kintsch, E. (2004). Summary street: Interactive computer support for writing. *Cogn. Instruct.* 22, 333–362. doi: 10.1207/s1532690xci2203_3

*Wang, Y. J., Shang, H. F., and Briody, P. (2013). Exploring the impact of using automated writing evaluation in english as a foreign language university students' writing. *Comput. Assist. Lang. Learn.* 26, 234–257. doi: 10.1080/09588221.2012.655300

Williams, C., and Beam, S. (2019). Technology and writing: Review of research. *Comput. Educ.* 128, 227–242. doi: 10.1016/j.compedu.2018.09.024

*Wilson, D. B. (2016). *Formulas Used by the Practical Meta-analysis Effect Size Calculator. Practical Meta-Analysis*. Unpublished manuscript: George Mason University. Available online at: https://mason.gmu.edu/~dwilsonb/downloads/esformulas.pdf

Wilson, J., and Czik, A. (2016). Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Comput. Educ.* 100, 94–109. doi: 10.1016/j.compedu.2016.05.004

*Wilson, J., and Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *J. Educ. Comput. Res.* 58, 87–125. doi: 10.1177/0735633119830764

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Front. Psychol.* 10, 3087. doi: 10.3389/fpsyg.2019.03087

*Zaini, A., and Mazdayasna, G. (2015). The impact of computer-based instruction on the development of EFL learners' writing skills. *J. Comput. Assist. Learn.* 31, 516–528. doi: 10.1111/jcal.12100

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 171. doi: 10.1186/s41239-019-0171-0

Zhai, N., and Ma, X. (2021). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Comput. Assist. Lang. Learn.* 2021, 1–26. doi: 10.1080/09588221.2021.1897019

Zhai, N., and Ma, X. (2022). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *J. Educ. Comput. Res.* 2022, 7356331221127300. doi: 10.1177/07356331221127300

Zhang, S. (2021). Review of automated writing evaluation systems. *J. China Comput. Assist. Lang. Learn.* 1, 170–176. doi: 10.1515/jccall-2021-2007

Zhu, M., Liu, O. L., and Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Comput. Educ.* 143, 103668. doi: 10.1016/j.compedu.2019.103668

Zhu, M., and Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *Eur. J. Psychol. Educ.* 30, 21–39. doi: 10.1007/s10212-014-0225-6

*References marked with an asterisk indicate studies included in the meta-analysis.

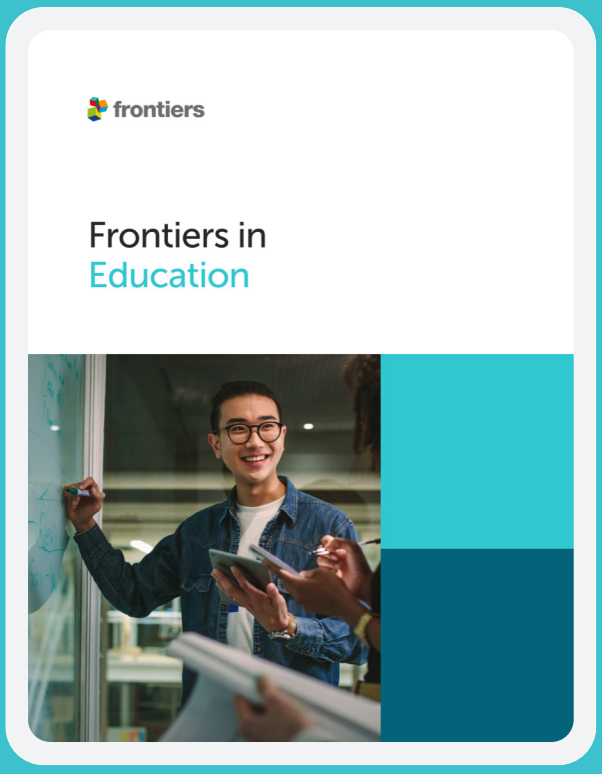
Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact