Computers & Education 210 (2024) 104967

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Impact of AI assistance on student agency

Ali Darvishi [a,*], Hassan Khosravi [b], Shazia Sadiq [b], Dragan Gašević [c], George Siemens [d]

[a] Business School, The University of Queensland, St Lucia, QLD, 4072, Australia
[b] School of Electrical Engineering and Computer Science, The University of Queensland, St Lucia, QLD, 4072, Australia
[c] Centre for Learning Analytics, Faculty of Information Technology, Monash University, Melbourne, VIC, 3800, Australia
[d] Centre for Change and Complexity in Learning, University of South Australia, Australia

ARTICLE INFO

Keywords:
AI in education
Student agency
Peer feedback
Educational technology

ABSTRACT

AI-powered learning technologies are increasingly being used to automate and scaffold learning activities (e.g., personalised reminders for completing tasks, automated real-time feedback for improving writing, or recommendations for when and what to study). While the prevailing view is that these technologies generally have a positive effect on student learning, their impact on students' agency and ability to self-regulate their learning is under-explored. Do students learn from the regular, detailed and personalised feedback provided by AI systems, and will they continue to exhibit similar behaviour in the absence of assistance? Or do they instead continue to rely on AI assistance without learning from it? To contribute to filling this research gap, we conducted a randomised controlled experiment that explored the impact of AI assistance on student agency in the context of peer feedback. With 1625 students across 10 courses, an experiment was conducted using peer review. During the initial four-week period, students were guided by AI features that utilised techniques such as rule-based suggestion detection, semantic similarity, and comparison with previous comments made by the reviewer to enhance their submissions if the feedback provided was deemed insufficiently detailed or general in nature. Over the following four weeks, students were divided into four different groups: control (AI) received prompts, (NR) received no prompts, (SR) received self-monitoring checklists in place of AI prompts, and (SAI) had access to both AI prompts and self-monitoring checklists. Results of the experiment suggest that students tended to rely on rather than learn from AI assistance. If AI assistance was removed, self-regulated strategies could help fill the gap but were not as effective as AI assistance. Results also showed that hybrid human-AI approaches that complement AI assistance with self-regulated strategies (SAI) were not more effective than AI assistance on its own. We conclude by discussing the broader benefits, challenges and implications of relying on AI assistance in relation to student agency in a world where we learn, live and work with AI.

1. Introduction

AI-powered educational technologies (AI-EdTech) are increasingly being used to automate and scaffold learning activities. For

* Corresponding author.
E-mail addresses: a.darvishi@uq.edu.au (A. Darvishi), h.khosravi@uq.edu.au (H. Khosravi), shazia@eecs.uq.edu.au (S. Sadiq), dragan.gasevic@monash.edu (D. Gašević), george.siemens@unisa.edu.au (G. Siemens).

https://doi.org/10.1016/j.compedu.2023.104967
Received 12 July 2023; Received in revised form 20 November 2023; Accepted 21 November 2023
Available online 30 November 2023

example, educational recommender systems suggest personalised learning resources (Bodily & Verbert, 2017), automated feedback systems provide real-time feedback on students' work (Cavalcanti et al., 2021; Deeva, Bogdanova, Serral, Snoeck, & De Weerdt, 2021), personalised nudging systems remind students of upcoming deadlines (Damgaard & Nielsen, 2018) and adaptive educational systems (Ma, Adesope, Nesbit, & Liu, 2014; VanLehn, 2011) tailor instructions and curriculum to the needs of individual students.

Recent advancements in AI-EdTech based on large language models (LLMs), exemplified by ChatGPT and others, have sent shockwaves through the higher education sector and reshaped its landscape, offering new possibilities and challenges (Milano, McGrane, & Leonelli, 2023; Jeon & Lee, 2023). LLMs have the potential to assist teachers and learners in various ways, including generating text for assignments and exams, answering questions, and providing teaching support; however, they have also raised critical questions about their integration into teaching and assessment practices (Yan et al., 2023; Alqahtani et al., 2023; Arachchige & Arosh, 2023). These models can exhibit unpredictable behaviours, and steering their actions remains a challenge (Bowman, 2023). Moreover, they do not inherently express values, emphasising the need for responsible usage (Bozkurt, 2023). As the use of LLMs becomes more prevalent in education, educators and institutions face critical decisions about how to integrate these tools effectively.

The use of AI-EdTech for automation and scaffolding of learning by design is intended to streamline learning where students are nudged towards activities they need to complete. However, this approach appears to contradict theories related to self-regulated learning (SRL), which is an essential competence and a lifelong learning skill Winne (2013) that involves individuals making informed decisions and developing the capacity to set learning goals, evaluate progress, reflect and learn from feedback throughout their lives (Council, 2006; HolonIQ, 2022; OECD, 2018; Winne, 2006; WorldBank, 2021). Accordingly, this article aims to conduct empirical studies that inform the impact of AI scaffold and nudges, facilitated by an LLM, on student learning and their agency and ability to regulate their own learning. Student agency, defined as the capacity for students to actively shape their learning experiences, make responsible decisions, and control their educational journey (OECD, 2018; Vaughn, 2020; Inouye, Lee, & Oldac, 2022), is central to this inquiry. The concept of agency aligns with dispositional and motivational dimensions, highlighting students' capacity to regulate actions, make informed choices, and navigate complex social contexts (Adie, Willis, & Van der Kleij, 2018; Code, 2020; Emirbayer & Mische, 1998). In this context, we pose three research questions to shed light on the impact of AI assistance on student agency and the potential for self-regulation strategies to complement or replace AI assistance. These research questions guide our investigation into the relationship between AI support and student learning and agency.

RQ1. *Do students learn from AI assistance, or do they tend to rely on them without learning?* Many AI-Edtech support student learning by providing personalised assistance in terms of detailed feedback, recommendations, nudges, etc. This begs the question of whether students actually learn from AI assistance or tend to rely on it without truly comprehending the material. A tangible example might be the use of spell checkers. Rimbar reports results from an experiment that illustrates learners use spell-checkers to correct errors; however, they have very little influence on correcting the errors on the cognitive level as learners tend to make the same mistakes in future exercises Rimbar (2017). This question is crucial as it helps determine whether AI is a valuable addition to the classroom or whether it might hinder student learning and agency.

RQ2. *Can AI assistance and nudges be replaced with self-regulation strategies after some time without compromising performance?* The second research question examines whether AI assistance can be used without compromising agency. Specifically, our aim is to explore whether it is possible for students to benefit from AI assistance and scaffolds for an initial period of time but then transition into a second phase where they rely less on AI and more on self-regulation strategies. This question can have interesting implications as it intersects technology and pedagogy and the development of learning processes that transition between the two.

RQ3. *Would complementing AI assistance with self-regulation strategies improve student performance?* Given that AI is becoming ubiquitous in our everyday lives, it is interesting to explore whether methods that utilise human-AI collaboration by using AI but also relying on student agency might be effective. Specifically, our aim is to study the effect of *complementing AI assistance* with self-regulation strategies to help students maintain agency and develop self-regulation. An existing limitation with many AI-Edtech is that they have been shown to have a weak connection to theoretical pedagogical perspectives Chen, Xie, Zou, and Hwang (2020); Zawacki-Richter, Marín, Bond, and Gouverneur (2019).

The relationship between agency and the surrounding structure is a central theme in various theoretical perspectives. While some view agency as independent, others consider it interdependent with the structural elements that shape educational experiences (Adie et al., 2018; Emirbayer & Mische, 1998). This dynamic perspective on student agency aligns with the understanding that agency is inherent in students' ability to effectively regulate and control their learning processes (Code, 2020). In our research, we draw upon Winne's model of SRL to emphasise the concept of agency in the context of peer feedback and AI assistance, where agency is defined as "an ontological concept that refers to, among other things, the capability to exercise choice in reference to preferences" (Winne, 2006, p. 8). Engaging students in peer feedback has been recognised as a beneficial approach that promotes higher-order learning and provides students with fast and detailed feedback on their work Zhu and Carless (2018). Providing lengthy comments in peer feedback has been suggested to also benefit reviewers Zhu and Carless (2018), as it may indicate greater effort put into reviews, contributing to learning and self-regulation (Baars, Wijnia, de Bruin, & Paas, 2020). However, there are some concerns and criticisms associated with the use of peer review, as feedback provided by students may be ineffective and of low quality (Bates, Galloway, Riise, & Homer, 2014; Denny, Luxton-Reilly, & Simon, 2009; Galloway & Burns, 2015; Tackett et al., 2018; Walsh, Harris, Denny, & Smith, 2018). Previous research has demonstrated that AI can be effectively used to prompt students who submitted feedback that is too short, generic or similar to previously provided feedback to revise and resubmit their review (Jia et al., 2021; Xiong, Litmaan, & Schunn, 2012). Our research here complements this prior research by conducting a between-subjects randomised controlled experiment in 10 courses from different disciplines to explore the impact of removing AI assistance and replacing or accompanying it with self-regulation strategies on

student agency in the context of peer feedback. For the first four weeks of the semester, the AI prompts were offered to all participants during the peer review process. The prompts aimed to promote checking the quality of comments provided and ask students to address issues detected with their textual feedback. In the next four weeks of the semester, participants were randomly assigned to one of four groups, including a control group with continued AI prompts (AI) (control group), a group without AI prompts (NR) (to respond to RQ1), a group with self-regulation strategies instead of AI prompts (SR) (to respond to RQ2), and a group with both self-regulation strategies and AI prompts (SAI) (to respond to RQ3).

Through this examination, we aim to contribute to a deeper understanding of student agency in the context of AI assistance and self-regulation strategies, shedding light on the pivotal role of students in peer review and feedback processes. Our research endeavours to address key questions surrounding the impact of AI on student learning and agency, the feasibility of transitioning from AI assistance to self-regulation, and the potential benefits of complementing AI assistance with self-regulation strategies. In doing so, we aim to inform pedagogically sound ways of integrating AI into education while preserving and enhancing student agency. In what follows, Section 2 presents related work on AI applications in education, self-regulated learning, and example use cases in peer review and feedback. Section 3 presents our methodological approach to responding to the questions. Using three research questions as a structure, Section 4 delivers the results of the experiment. Finally, the ramifications of incorporating the AI-assistance into an educational system, as well as potential advantages and drawbacks, are discussed in Section 5. We also propose future research directions to address present restrictions.

2. Related work

AI in Education and Student Agency. In recent years, there has been considerable growth in educational technologies and a massive increase in learners' data and digital traces, presenting new opportunities for the application of AI in education (Gašević, Siemens, & Sadiq, 2023). The use of AI in education aims to support learners through personalised learning experiences, adaptive instructions, intelligent tutoring systems, immersive learning technologies, and automatic content creation (du Boulay, Mitrović, & Yacef, 2023; Hwang, Xie, Wah, & Gašević, 2020). Adaptive learning systems, for instance, collect data on students' interactions with various learning activities and adjust the content accordingly to provide personalised learning experiences. However, there is a concern that complete automation of the learning process by AI might limit students' development of agency and essential skills for the future, as studies have shown that AI systems, despite their potential benefits, can risk violating social boundaries, raise concerns about responsibility, agency, and surveillance issues, and challenge traditional notions of ethics and authenticity in education (Celik, Dindar, Muukkonen, & Järvelä, 2022; Darvishi, Khosravi, Sadiq, & Weber, 2022; Fyfe, 2022; Molenaar, 2022a; Seo, Tang, Roll, Fels, & Yoon, 2021). One way AI can support students' agency is through the implementation of learner models, personalised nudges, and recommender systems (Abdi, Khosravi, Sadiq, & Darvishi, 2021; Afzaal et al., 2021; Bodily & Verbert, 2017). These AI applications may assist students in developing self-regulation skills and making informed decisions (Suh, 2019). For instance, nudging systems can remind students of upcoming deadlines (Damgaard & Nielsen, 2018) and AI-powered tools can provide automated personalised feedback on student work or suggest improvements in their writing (Wambsganss, Janson, & Leimeister, 2022).

Self-Regulated Learning. On the other hand, self-regulated learning (SRL) remains critical for academic success, yet many students struggle without sufficient instructional support. Advances in technology and learning analytics have opened up opportunities to enhance SRL by improving students' metacognitive and cognitive processes (Baars et al., 2020). Customised learning environments, ranging from fixed to dynamic adaptive scaffolds, enable educators to effectively implement SRL interventions and provide metacognitive support (Azevedo & Hadwin, 2005). AI initiatives are also aimed at enhancing the effectiveness of SRL interventions in educational settings (Dawson, Joksimovic, Mills, Gašević, & Siemens, 2023; Hilpert, Greene, & Bernacki, 2023). For instance, they have been used to measure and augment SRL processes, provide data-driven feedback and action recommendations, and consider moderating factors like gender differences and need satisfaction (Järvelä, Nguyen, & Hadwin, 2023; Jin, Im, Yoo, Roll, & Seo, 2023; Heikkinen, Saqr, Malmberg, & Tedre, 2023; Afzaal et al., 2021; Xia, Chiu, & Chai, 2023). However, incorporating SRL techniques into educational technologies and establishing empirical evidence of their effectiveness are ongoing processes, demanding further research to identify best practices and gather robust evidence on their impact on student learning outcomes (Broadbent & Poon, 2015). Nonetheless, student agency, encompassing self-regulation, informed decision-making, and ownership of learning, has been linked to positive educational outcomes across various domains (Stenalt & Lassesen, 2022). One recognised and effective strategy for self-regulation is self-monitoring (Zimmerman & Paulsen, 1995). A number of studies have shown its effectiveness in various contexts, such as math homework, where it exhibited a positive linear trend in self-regulation (Schmitz & Perels, 2011). In online learning environments, self-monitoring techniques have been used to improve note-taking and enhance achievement (Kauffman, Zhao, & Yang, 2011). Moreover, in Massive Open Online Courses (MOOCs), self-monitoring has facilitated self-directed learning (Zhu & Bonk, 2019). Shibani (2019) further elaborates that self-monitoring/assessment improves students' evaluative judgment, empowering them to self-regulate their work and fostering sustainable learning. Notably, self-monitoring has also been incorporated within the peer review process to enhance the quality of peer feedback. For instance, Kulkarni, Bernstein, and Klemmer (2015) proposed the use of scaffold comments within PeerStudio, allowing students to re-review their peer feedback before its final submission.

Peer Review and Feedback. Additionally, initiatives have been undertaken to pinpoint the features of high-quality feedback, such as its length, specificity or localisation of the problem, scope, alignment to content, suggestions for improvement, and the use of constructive language (Nelson & Schunn, 2009; Xiong et al., 2012; Kovanović et al., 2016; Cavalcanti et al., 2020; Darvishi, Khosravi, & Sadiq, 2020; Jensen, Bearman, & Boud, 2021; Zong et al., 2021). In recent years, natural language processing (NLP) approaches have gained popularity in different educational contexts, notably in automatically analysing students' work and offering personalised automated feedback (Shibani, 2019). For example, Jia et al., 2021 assessed the quality of peer review comments using the BERT and

DistilBERT language representation models (Sanh, Debut, Chaumond, & Wolf, 2019). Xiong et al. (2012) presented a mix of NLP approaches and machine learning to detect the lack of significant features in peer review. Krause et al. (2017) proposed an NLP technique for automatically analysing feedback language and extracting features such as specificity and sentiment. In most of these investigations, user trials have been undertaken to demonstrate the value of NLP techniques in raising the standard of student-generated feedback. Large language models, such as ChatGPT, hold transformative potential in education, enabling improvements in academic writing assistance, assisting educators in various aspects of pedagogy, enhancing productivity, generating detailed feedback, and reshaping the peer review process while also raising noteworthy considerations related to ethics, biases, and transparency (Cooper, 2023; Donker, 2023; Li, Patel, & Du, 2023; Meyer et al., 2023). For instance, Hosseini and Horbach (2023) examined how these models could impact the peer review process, highlighting their potential to enhance reviewer and editor roles, albeit with concerns about biases and confidentiality. Bauer et al. (2023) proposed a framework for integrating NLP into peer feedback processes, offering potential avenues to innovate digital learning environments. Dai et al. (2023) found that ChatGPT could provide detailed and effective feedback to students, showcasing its potential to improve the feedback generation process. And Watkins (2023) emphasised the importance of developing norms and guidelines for researchers and peer reviewers using large language models in scientific investigations, aligning with AI ethics principles to ensure ethical and responsible use.

The majority of these initiatives, however, have concentrated on certain elements or strategies in an effort to improve students' capacity to offer helpful peer feedback. Addressing the multifaceted nature of effective feedback, the importance of synergistic interactions and the necessity to satisfy multiple conditions has been underlined (Darvishi, Khosravi, Sadiq, & Gašević, 2022; Henderson et al., 2019). Molenaar (2022b) proposed a hybrid model combining human and artificial intelligence to leverage their strengths and improve teaching and learning outcomes. Yet, understanding the long-term effects of using AI assistance to scaffold learning on student agency and their ability to regulate their work independently is an under-explored area (Moses, Rylak, Reader, Hertz, & Ogden, 2020).

Research Gap and Contribution. This paper presents the findings of a study that investigated the impact of AI assistance on student agency within the context of peer feedback. In the realm of feedback, it is crucial to develop a more comprehensive understanding of student agency, specifically regarding the processes of seeking, receiving, creating, and acting on feedback information, which are often overlooked despite their significance in implementing effective learner-centred feedback approaches (Nieminen, Tai, Boud, & Henderson, 2022). The study aimed to examine whether the use of AI, employed to monitor the quality of peer feedback, assists students in developing the ability to provide effective feedback without relying on AI assistance. Furthermore, we explored the potential of self-regulation strategies as alternatives or complements to AI assistance without compromising performance levels.

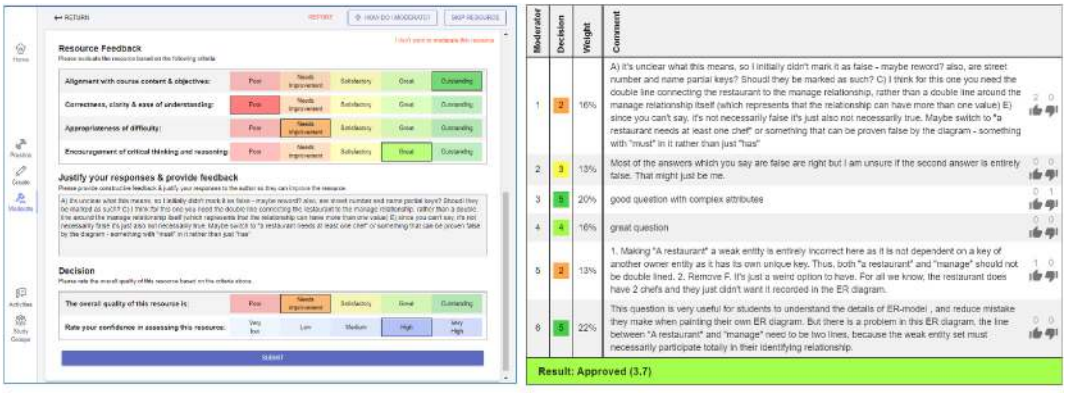
3. Methods

In this section, we first introduce the peer review process in RiPPLE, the tool used in our study. Then, we describe the study design, including the experimental groups, participants, and measures used for analysis.

3.1. Peer review in RiPPLE

RiPPLE is an adaptive education system that uses learnersourcing (Khosravi, Demartini, Sadiq, & Gasevic, 2021; Kim et al., 2014) by involving students in creating and evaluating learning resources. A student-created resource must go through peer review. When a student is available to review a resource, RiPPLE presents them with an unmoderated resource. Resources are typically selected based on a first-in-first-out queue, but factors such as concurrency and conflict of interest may impact the selection.

Fig. 1a presents the peer review interface used in RiPPLE. Moderators rate resources using a 4-item rubric that evaluates the alignment, correctness, difficulty level, and critical thinking. They are requested to provide feedback to the author and their confidence in the rating. If an instructor is available, the system uses spot-checking algorithms to identify resources that could benefit more from



(a) Moderation Rubric-only interface (b) Moderation outcome and provided feedback

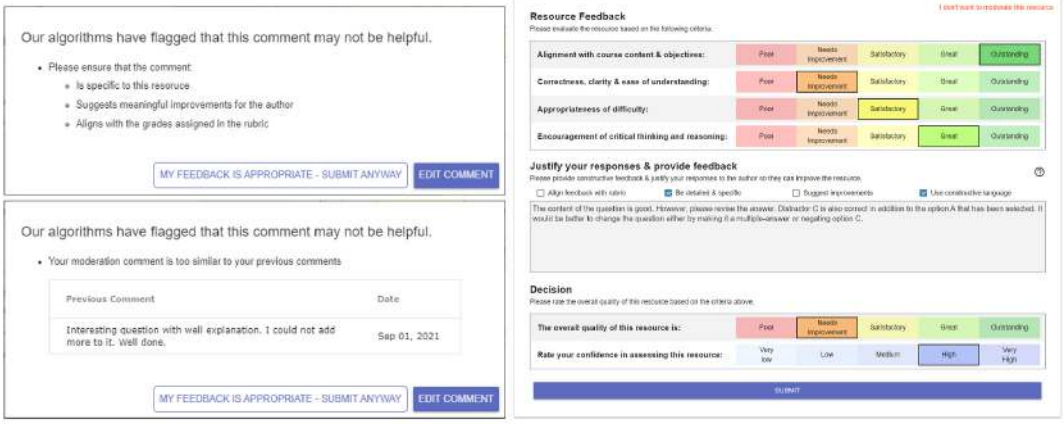
Fig. 1. Interfaces of peer review processes in RiPPLE.

an expert review (Darvishi, Khosravi, & Sadiq, 2021). Reviews from instructors are final and determine the quality of the resource. The system uses consensus algorithms to update the resource's status and reliability ratings of the author and student moderators (Darvishi, Khosravi, Rahimi, Sadiq, & Gašević, 2022). Approved resources go into the repository, while rejected ones can be resubmitted after updates. Fig. 1b shows how peer review outcomes and feedback are shared. Instructors can view student moderator names, ratings, confidence levels, and feedback. The author and student moderators can see the ratings and feedback but not the identity or reliability ratings of other moderators. They can also mark feedback as helpful/unhelpful by providing likes/dislikes.

3.1.1. AI-assistance

Recent advancements in NLP have been driven by the development of large language models, such as GPT-3, RoBERTa, ELECTRA, and ALBERT (Brown et al., 2020; Clark, Luong, Le, & Manning, 2020; Lan et al., 2019; Liu et al., 2019). With their impressive parameter sizes and fine-tuning capabilities, these transformer-based LLMs have propelled NLP to new heights of performance in tasks like question-answering, text completion, and language translation. The success of LLMs has found applications in various domains, including chatbots and automated assistants, where their ability to understand and respond to human language has dramatically improved. Additionally, these models have demonstrated their effectiveness in unsupervised pre-training on large text corpora, leading to the development of powerful embedding models like SBERT (Sentence-BERT: Bidirectional Encoder Representations from Transformers) (Reimers & Gurevych, 2019). SBERT, leveraging BERT's encoding capabilities (Devlin, Chang, Lee, & Toutanova, 2018), has exhibited superior performance in text similarity tasks, making it a practical tool for our application. In the context of our study, we have harnessed the developments in NLP and large language models to implement AI assistance for peer feedback in RiPPLE. This AI assistance is underpinned by a set of automated quality control functions designed to evaluate and enhance the quality of student-submitted feedback. These functions encompass: (1) *Suggestion Detection*: One of our functions is dedicated to identifying suggestions within the submitted feedback, a rule-based technique inspired by the work of (Negi, Asooja, Mehrotra, & Buitelaar, 2016). It scans the text to determine if any suggestions are present. If none are detected, the system triggers prompts to encourage students to provide constructive suggestions. (2) *Relatedness Score*: To measure how closely the submitted feedback aligns with the context of the learning resource, we employ SBERT, a state-of-the-art embedding model (Reimers & Gurevych, 2019). SBERT leverages the power of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and calculates a cosine similarity score between the feedback and the resource context. This score provides a measure of their semantic relatedness, ranging from −1 to 1. When the Relatedness Score falls below a certain threshold, the system prompts students to ensure that their comments are specific and contextually relevant to the resource. This encourages more meaningful and applicable feedback. (3) *Similarity Score*: Our third function assesses the similarity between the submitted text and previous comments made by the moderator, utilising the GLEU (Google's biLingual Evaluation Understudy) measure (Wu et al., 2016). By analysing n-grams, this function helps prevent repetitive and generic comments.

The primary objectives of these functions are to ensure that feedback is not only relevant but also original and constructive. Moreover, they serve to reduce the likelihood of students trying to manipulate the system by submitting repetitive comments. The AI-assisted approach guides students through the feedback process, offering prompts as illustrated in Fig. 2a, highlighting potential issues with the provided feedback and asking students to revise their comments where the AI identifies room for improvement, thus fostering a culture of continuous improvement in feedback provision. When presented with these prompts, students have the option of acting on their own to either edit their feedback independently in response to the suggestions or confirm that their input is appropriate as is. In essence, the incorporation of these AI-driven functions is twofold in purpose: firstly, it introduces an element of oversight and accountability, encouraging students to provide more thoughtful and valuable feedback, thereby mitigating the issue of poor-quality contributions. Secondly, the prompts provide clarity regarding why the system suggests improvements, aiding students in understanding the rationale behind the AI assistance.



(a) Examples of AI-assistance prompts. (b) Peer review interface with a self-monitoring checklist.

Fig. 2. Complementary peer review interfaces include (a) AI-assistance prompts and (b) a self-monitoring checklist.

3.1.2. Self-monitoring checklist

First, a set of training materials was developed for students to write constructive and effective feedback to enhance feedback quality. The training covers four key criteria: align with rubrics, be detailed and specific, suggest improvements, and use constructive language. Examples of good and bad practices are included. Then, acknowledging the effect of externally-facilitated monitoring on self-regulated learning (Gašević, Adesope, Joksimović, & Kovanović, 2015; Gašević, Mirriahi, Dawson, & Joksimović, 2017), a self-monitoring checklist was incorporated to scaffold an essential SRL process of activating prior knowledge (Azevedo, Moos, Greene, Winters, & Cromley, 2008). This checklist supports the guidelines in the training material and empowers students to monitor their adherence to the feedback criteria. As shown in Fig. 2b, the training materials can be accessed through a (?) button, and the checklist is displayed on top of the feedback input box. This serves to help students enhance their feedback literacy and foster agency and self-regulation in their learning.

3.2. Experiment design

The study was a randomised controlled field experiment, with the peer review condition as the independent variable (one control group and three experimental groups). The dependent variables were the quality of the comments or feedback measured by different metrics, including the rate of flagged reviews (comments needed revision), similarity (rate of generic comments), relatedness (comment-resource pair), comment length, time spent on review, and likes/helpfulness. The experiment took place over eight weeks, with AI prompts provided to all students in the first four weeks, and then students were randomly divided into four groups in the second four weeks.

- **Control group (AI):** AI-assisted group –prompts remained in the peer review interface.
- **Experiment 1 (NR):** Not Receiving assistance group – the AI assistance prompts were removed from the peer review interface. A comparison of the results of the AI and NR groups, reported in Section 4.1, is used for responding to RQ1.
- **Experiment 2 (SR):** Self-monitoring checklist group – the AI assistance prompts were replaced with a self-monitoring checklist and a set of guidelines in the peer review interface. A comparison of the results of the AI and SR groups, reported in Section 4.2, is used for responding to RQ2.
- **Experiment 3 (SAI):** Self-monitoring and AI-assisted group – the AI assistance prompts were complemented with the self-monitoring checklist in the peer review interface. A comparison of the results of the AI and SAI groups, reported in Section 4.3, is used for responding to RQ3.

3.2.1. Participants

In this study, we collected data from undergraduate students enrolled in 10 courses that used RiPPLE in the second semester of 2020. These courses represented various disciplines, such as Humanities & Social Sciences, Health & Behavioural Sciences, Medicine, Business, and Engineering. Only participants who consented to have their data used in the study were included. During the first 4 weeks of the experiment, all participants received AI prompts during peer review. In the second 4 weeks, 1625 students submitted 11,243 peer reviews on 3573 resources across the 10 courses. Table 1 summarises the number of students and the number of peer reviews for each group during the second 4 weeks of the experiment.

3.2.2. Measures

We have formulated three research questions to investigate the impact of AI assistance on student agency and the potential for self-regulation strategies to replace or complement AI assistance. In this section, we present a set of measures that will be used across all research questions as our means to assess the quality of student feedback, thereby allowing us to capture various dimensions of student agency aligned with different facets of COPES in Winne's model Winne (2013). In the context of SRL, it is suggested that agency is evident through five key elements referred to as COPES, which encompass conditions, operations, products, evaluation, and standards, as outlined in Winne's model (Greene & Azevedo, 2007). In our study, we have operationalised agency by assessing metrics like comment length, similarity, relatedness, number of likes, and more, aligning these measures with different facets of COPES. These metrics illuminate learners' choices related to the standards applied to their products (peer feedback in this context), highlighting their capacity to deliver contextually meaningful and non-generic feedback, thus emphasising their independent and self-regulation abilities in the learning process. By employing multiple measures, we aim to shed light on the different aspects of student agency, the

Table 1
Overview of the collected data of student peer review for the number of students and peer reviews in each experiment group, and the total number of resources under review.

Peer review condition	# Students	# peer reviews	# Resources
AI	396	2725	
NR	409	2757	
SR	402	2894	3573
SAI	418	2867	
Total	1625	11,243	

effectiveness of AI assistance and self-regulation strategies, and gauge the extent to which students take responsible action and exhibit agency by independent decision-making when AI assistance is absent. It allows us to capture various dimensions of student feedback and assess their engagement and learning outcomes. Each measure provides a distinct perspective on student agency, contributing to a more holistic evaluation and enhancing the validity and reliability of our findings, as it helps to overcome potential biases or limitations that may arise from relying on a single measure.

- **Rate of Flagged Reviews** serves as an indicator of the extent to which students' comments require revision based on the quality control functions. When comments are flagged, it suggests that the reviews may lack specific details or suggestions related to the assessed resource. This measure aligns with the objective of understanding whether students learn from AI assistance or tend to rely on it without actively engaging in the learning process. It also shows students' acceptance of the system and engagement with the AI suggestions, which plays a significant role in their agency and success Tsai (2015). The rate of flagged reviews serves as a valuable indicator of student agency in peer review, reflecting their active engagement in refining feedback and independent commitment to improving its quality. In essence, monitoring and controlling flagged reviews represent key aspects of student agency, demonstrating their responsible actions and decision-making in the absence of external guidance.
- **Similarity Score** represents the average similarity between a student's comments and their previous comments, as determined by GLEU. Essentially, it assesses the extent to which students incorporate their own previous comments into their current reviews. It serves as a valuable indicator of metacognitive processes, offering insight into students' ability to self-regulate and fine-tune their feedback over time. When the similarity score is lower, it signifies that students are actively engaged in the process of avoiding generic and repetitive feedback, opting instead to provide pertinent comments. Within the context of our research, this metric is not just a numerical value; it provides a window into learners' decision-making regarding the standards they apply to their product–peer feedback. It demonstrates their dedication to steering clear of generic commentary and underscores their commitment to delivering thoughtful feedback. Importantly, this capability to regulate feedback aligns seamlessly with the broader concept of agency, underscoring their ability to autonomously govern their learning experiences.
- **Relatedness Score** reflects the average relatedness between a student's comments and the resources under review, as determined by SBERT. This measure examines the extent to which students address the specific context and content of the assessed resources in their feedback. Research has shown that being specific and detailed in feedback increases implementation and is more valuable than generic praises or criticisms (Henderson et al., 2019; Nelson & Schunn, 2009). Therefore, it not only underscores students' capacity to make deliberate choices concerning the quality and specificity of their comments but also represents their capability to consider the context and content when offering feedback. This metric essentially reflects the dynamic interaction between conditions (such as their knowledge, motivation, and the resources available) and operations (the cognitive processes they engage in) as students make conscious decisions about the relevance and appropriateness of their comments.
- **Length** shows the average number of words in comments made by a student. While longer comments do not guarantee higher quality, research suggests a positive association between feedback quality and comment length (Zong et al., 2021). The significance of this metric in evaluating student agency is further demonstrated by the finding that the quantity of words is a leading predictor of quality in terms of cognitive presence (Kovanović et al., 2016; Page & Petersen, 1995). This metric aligns with the standards component, shedding light on students' criteria for effective feedback.
- **Time:** captures the average time in seconds spent on writing comments by a student, serving as an indicator of their engagement and dedication. Research has shown that investing more time in completing a review is associated with increased effort, which in turn contributes to learning and self-regulation processes (Baars et al., 2020). Therefore, the time measure can serve as an indicator of students' commitment to the feedback task and their potential for self-regulation. It assesses the impact of different strategies on performance, as it provides insights into students' level of involvement and their willingness to invest time in the feedback process. Time spent corresponds to the operations element of Winne's SRL model, signifying the cognitive processes engaged in activities like searching, monitoring, assembling, etc. It illustrates students' active engagement and commitment to conducting review tasks.
- **Rate of Likes (Helpfulness):** represents the likelihood of a student's comments receiving votes of helpfulness from other reviewers. It is a commonly adopted metric to evaluate the quality of peer assessment (Misiejuk & Wasson, 2021). This measure reflects the perceived value and impact of students' comments, providing insights into the effectiveness of different strategies in facilitating feedback that is well-received and considered valuable. It contributes to establishing standards for the quality and impact of feedback, highlighting its role in evaluating the success of the review process.

3.2.3. Analysis

The study included one experiment with three treatments: Experiment 1 (NR) representing the Not Receiving assistance group, Experiment 2 (SR) representing the Self-monitoring checklist group, and Experiment 3 (SAI) representing the Self-monitoring and AI-assisted group. Each treatment was designed to address a specific research question, with Experiment 1 examining the impact of removing AI assistance (RQ1), Experiment 2 investigating the potential replacement of AI assistance with self-regulation strategies (RQ2), and Experiment 3 exploring the effects of complementing AI assistance with self-regulation strategies on student performance (RQ3). This approach enabled a focused and extensive analysis of the various metrics, including the flags on quality of comments, similarity, relatedness, comment length, time spent on review, and likes/helpfulness, in relation to each research question and treatment. The statistical analysis was conducted using R and SPSS software for visualisation, data analysis, and statistical tests. One-way ANOVA was employed to determine significant differences between peer review conditions, followed by Tukey's HSD (honestly significant difference) post-hoc tests to identify specific pairwise differences. This approach facilitated a detailed understanding of nuanced effects for each peer review condition, allowing examinations of the impact of different conditions on the various measures.

4. Results

Appendix A summarises student data collected during the initial four weeks of exposure to AI prompts. The results of one-way ANOVA tests, as outlined in the appendix, reveal no significant difference between students in the four groups, underscoring the uniformity in their peer review behaviours during this period. Fig. 3 compares the students' peer review behaviour during the second four weeks between the control group (AI) and the students in the three experiment groups (NR, SR, and SAI) in terms of the six different measures mentioned in Section 3.2.2. The boxplots show data distributions as a five-number summary: minimum, first quartile, median, third quartile, and maximum. The highlighted points indicate the mean. Also, the violin plots were included to display the overall distribution of the data using density curves. Fig. 3a shows the rate of reviews flagged by the system for each user in the control group and three experimental groups. A one-way ANOVA revealed that there was a statistically significant difference in flag rates between at least two groups ($F(3, 1621) = 39.63, p < 0.001, \eta^2 = 0.068$). Fig. 3b shows the average similarity score with previous comments for each user in the four groups. A one-way ANOVA revealed that there was a statistically significant difference in the average similarity score between at least two groups ($F(3, 1621) = 9.65, p =< 0.001, \eta^2 = 0.018$). Fig. 3c shows the average relatedness score of comments and resources under review for each user in the four groups. An ANOVA with relatedness score as the dependent variable and review condition as the independent variable revealed a significant difference between conditions ($F(3, 1621) = 12.24, p =< 0.001, \eta^2 = 0.022$). Fig. 3d shows the average length of comments in words for each user in the four groups. A one-way ANOVA revealed that there was a statistically significant difference in comments' length between at least two groups ($F(3, 1621) = 14.14, p < 0.001, \eta^2 = 0.026$). Fig. 3e shows the average time spent providing peer reviews for each user in the four groups. A one-way ANOVA revealed no statistically significant difference in the average time spent between conditions ($F(3, 1621) = 0.44, p = 0.728, \eta^2 = 0.001$). Fig. 3f shows the rate of reviews that received a like/helpfulness for each user in the four groups. A one-way ANOVA revealed that there was not a statistically significant difference in the like rates between conditions ($F(3, 1621) = 1.07, p = 0.359, \eta^2 = 0.002$). In the rest of this section, we compare each of the experimental groups with the control group to answer our three proposed research questions.

4.1. RQ1: removing AI-assistance (AI vs NR)

Table 2 compares the peer review behaviour of students in the control group (AI) with the students in the first experiment group (NR) in terms of the six different measures mentioned in Section 3.2.2. The table includes various metrics, each providing insights into the differences between the two groups. 'Mean' and 'SD' represent the average and standard deviation, respectively. The '95% CI for Mean Difference' indicates the range within which the true mean difference is likely to fall. 't' denotes the t-statistic, measuring the size of the difference relative to the variability in the data. 'Cohen's d' indicates the standardised effect size. 'p_{tukey}' represents the p-value from Tukey's HSD post-hoc test, indicating the probability of observing a difference as extreme as the one computed, assuming the null hypothesis is true. Tukey's HSD test found that the mean value of flagged review percentage was significantly different between AI and NR ($p < 0.001$, 95% $CI = [-0.22, -0.12]$). As shown, the students in the NR group received significantly higher rates of flags ($M = 0.31$, $SD = 0.31$) than the students in the AI group ($M = 0.14$, $SD = 0.23$), $t = -8.70$, and *Cohen's* $d = -0.61$ shows an effect size of medium to large. The second row shows the average similarity score with previous comments for each user in AI and NR. The comments of students in the NR group were significantly more similar ($M = 0.32$, $SD = 0.17$) to their previous comments than those of the students in the AI group ($M = 0.28$, $SD = 0.12$), $t = -4.22, p < 0.001$, 95% $CI[-0.07, -0.02]$, and *Cohen's* $d = -0.30$ shows a small

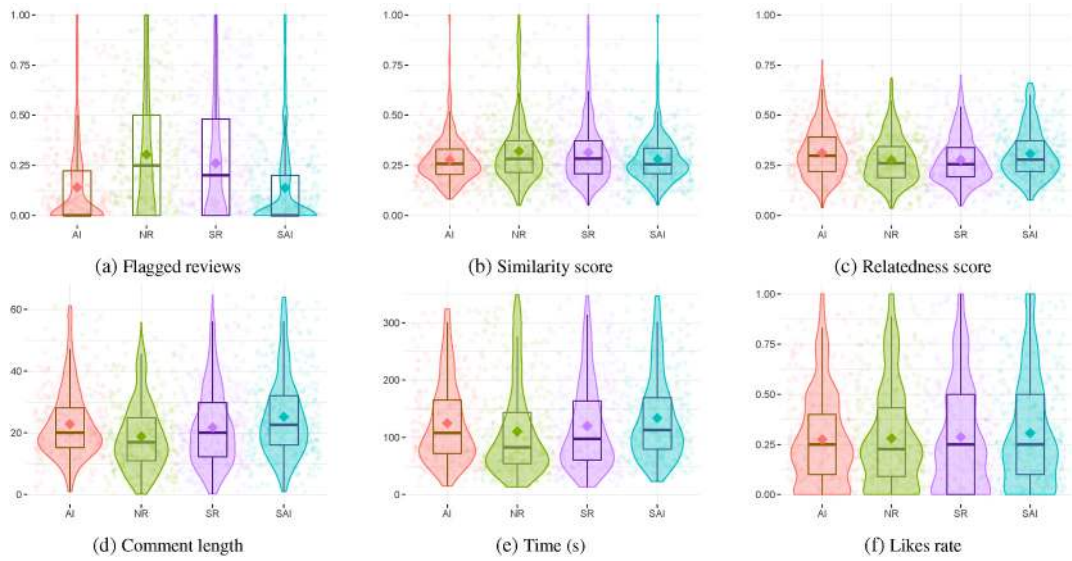


Fig. 3. Comparing the control group (AI) with the three experimental groups (NR, SR, and SAI) using six measures.

Table 2
Comparison between the control group (AI) and the first experimental group (NR), removing AI-assistance prompts.

AI vs NR	AI (N = 396)		NR (N = 409)		95% CI for Mean Difference		t	Cohen's d	p_{tukey}
Measure	Mean	SD	Mean	SD	Lower	Upper			
Flag Rate	0.14	0.23	0.31	0.31	−0.22	−0.12	−8.70	−0.61	<.001
Similarity Score	0.28	0.12	0.32	0.17	−0.07	−0.02	−4.22	−0.30	<.001
Relatedness Score	0.31	0.13	0.27	0.12	0.02	0.06	4.70	0.33	<.001
Comment length	23.38	12.41	19.43	13.15	1.43	6.49	4.03	0.28	<.001
Time	265.44	988.78	219.36	866.23	−103.88	196.03	0.79	0.06	0.859
Like Rate	0.28	0.26	0.28	0.27	−0.05	0.05	−0.01	-8×10^{-4}	1.000

to medium effect size. Tukey's HSD test on the average relatedness score of comments and resources under review for each user in AI and NR showed that the comments of students in the NR group were significantly less related ($M = 0.27$, $SD = 0.12$) to the resources under review than those of the students in the AI group ($M = 0.31$, $SD = 0.13$), $t = 4.70$, $p < 0.001$, 95% CI[0.02, 0.06], with a small to medium effect size (*Cohen's d* = 0.33). The comparison of the average length of comments between AI and NR groups showed that the comments of students in the NR group were significantly shorter ($M = 19.43$, $SD = 13.15$) than those of the students in the AI group ($M = 23.38$, $SD = 12.41$), $t = 4.03$, $p < 0.001$, 95% CI[1.43, 6.49], and the *Cohen's d* effect size is 0.28. Tukey's HSD test did not find a significant difference in time spent on reviews in the NR group compared to the AI group (NR: $M = 219.36$ & $SD = 866.23$, AI: $M = 265.44$ & $SD = 988.78$; $t = 0.79$, $p = 0.859$, 95% CI[−103.88, 196.03], *Cohen's d* = 0.06), and also in like rates (NR: $M = 0.28$ & $SD = 0.27$, AI: $M = 0.28$ & $SD = 0.26$; $t = -0.01$, $p = 1.000$, 95% CI[−0.05, 0.05], *Cohen's d* = -8×10^{-4}).

4.2. RQ2: replacing AI-assistance with self-monitoring checklist (AI vs SR)

Table 3 compares the peer review behaviour of students in the control group (AI) with the students in the second experiment group (SR) in terms of the six different measures. Tukey's HSD test indicated significant differences ($p < 0.001$) between AI and SR in terms of the rate of flagged reviews ($p < 0.001$, 95% $CI = [-0.17, -0.07]$). The flag rate of students in the SR group was significantly higher ($M = 0.26$, $SD = 0.30$) than that of the students in the AI group ($M = 0.14$, $SD = 0.23$), $t = -6.36$, $p < 0.001$, and *Cohen's d* = −0.45 showed a medium effect size. Tukey's HSD test also revealed a significant difference in the similarity score between AI and NR ($p = 0.008$, 95% $CI = [-0.06, -0.01]$). The comments' similarity for students in the SR group ($M = 0.31$, $SD = 0.15$) was significantly higher than that of the students in the AI group ($M = 0.28$, $SD = 0.12$), $t = -3.20$, and *Cohen's d* = −0.23 showed a small effect size. Tukey's HSD test found a significant difference in the relatedness score between AI and SR ($p < 0.001$, 95% $CI = [0.02, 0.06]$). The comments of students in the SR group were significantly less related ($M = 0.28$, $SD = 0.12$) to the resources under review than those of the students in the AI group ($M = 0.31$, $SD = 0.13$), $t = 4.26$, with a small to medium effect size (*Cohen's d* = 0.30). The post hoc Tukey's HSD test found no statistically significant difference between AI and SR groups in terms of the length of comments (SR: $M = 23.06$ & $SD = 15.26$, AI: $M = 23.38$ & $SD = 12.41$; $t = 0.32$, $p = 0.988$, 95% CI [−2.22, 2.86], *Cohen's d* = 0.02), time (SR: $M = 220.23$ & $SD = 481.87$, AI: $M = 265.44$ & $SD = 988.78$; $t = 0.77$, $p = 0.87$, 95% CI[−105.39, 195.81], *Cohen's d* = 0.06), and like rate (SR: $M = 0.31$ & $SD = 0.30$, AI: $M = 0.28$ & $SD = 0.26$; $t = -0.55$, $p = 0.947$, 95% CI[−0.06, 0.04], *Cohen's d* = −0.04).

4.3. RQ3: accompanying AI-assistance with self-monitoring checklist (AI vs SAI)

Table 4 compares the peer review behaviour of students in the control group (AI) with the students in the third experiment group (SAI) across the various measures. The flagged review percentage did not show a significant difference between the AI group ($M = 0.14$, $SD = 0.23$) and the SAI group ($M = 0.14$, $SD = 0.24$), $t = -0.10$, $p = 1.000$, 95% CI[−0.05, 0.05], *Cohen's d* = −0.01. Similarly, the similarity score between students' comments and previous comments did not significantly differ between the AI group ($M = 0.28$, $SD = 0.12$) and the SAI group ($M = 0.28$, $SD = 0.12$), $t = 0.00$, $p = 1.000$, 95% CI[−0.03, 0.03], *Cohen's d* = 7×10^{-5}. The comparison

Table 3
Comparison between the control group (AI) and the second experimental group (SR), replacing AI-assistance prompts with the self-monitoring checklist.

AI vs SR	AI (N = 396)		SR (N = 402)		95% CI for Mean Difference		t	Cohen's d	p_{tukey}
Measure	Mean	SD	Mean	SD	Lower	Upper			
Flag Rate	0.14	0.23	0.26	0.30	−0.17	−0.07	−6.36	−0.45	<.001
Similarity Score	0.28	0.12	0.31	0.15	−0.06	−0.01	−3.20	−0.23	0.008
Relatedness Score	0.31	0.13	0.28	0.12	0.02	0.06	4.26	0.30	<.001
Comment length	23.38	12.41	23.06	15.26	−2.22	2.86	0.32	0.02	0.988
Time	265.44	988.78	220.23	481.87	−105.39	195.81	0.77	0.06	0.867
Like Rate	0.28	0.26	0.31	0.30	−0.06	0.04	−0.55	−0.04	0.947

of the relatedness score of comments to resources under review also did not yield a significant difference between the AI group ($M = 0.31$, $SD = 0.13$) and the SAI group ($M = 0.31$, $SD = 0.13$), $t = 0.50$, $p = 0.959$, 95% CI[−0.02, 0.03], with a small effect size (*Cohen's d* = 0.04). However, the comparison of the average length of comments between the AI group ($M = 23.38$, $SD = 12.41$) and the SAI group ($M = 25.67$, $SD = 14.75$) showed a marginal difference, but it was not statistically significant ($t = -2.34$, $p = 0.090$, 95% CI[−4.80, 0.23]), with a small effect size (*Cohen's d* = −0.16). Additionally, there were no significant differences observed in the time spent on reviews ($t = -0.04$, $p = 1.000$, 95% CI[−151.49, 146.83], *Cohen's d* = −0.003) and like rates ($t = -1.55$, $p = 0.408$, 95% CI[−0.08, 0.02], *Cohen's d* = −0.11) between the SAI and AI groups.

5. Discussion

Our study showed that the integration of AI in learning environments could impact students' agency to take control of their own learning. Through a randomised controlled experiment, we found that while students can effectively self-regulate their learning with the aid of AI, removing this support would significantly change their performance. While the hybrid human-AI approach in the SAI group had the highest average performance among other groups, its improvement was not significant compared to the AI-only approach. These findings suggest that as AI becomes more prevalent in education, it is important to consider the role it plays in shaping student agency. Further research is needed to comprehend the complex relationship between AI and agency in learning contexts. There is a rising awareness of the hazards of outsourcing self-regulated learning to technology, which may impede students' cognitive and metacognitive growth (Molenaar, 2022b).

In response to RQ1, the findings of our experiment in Table 2 suggest that the AI prompts played a significant role in maintaining the quality of students' feedback. This is demonstrated by the higher flag rate in the no-support group (NR) after the AI support was removed, as well as the higher similarity score of the NR group's comments in comparison to the AI group. These results suggest that the AI prompts helped students avoid providing generic feedback and encouraged them to consider the unique aspects of the resource under review. In addition, the lower relatedness score of comments to the resource in the NR group after the removal of AI support suggests that the prompts may have helped students provide more specific and targeted feedback. This is consistent with the research of Nelson and Schunn (2009); Henderson et al. (2019), who found that feedback quality is strongly associated with specificity. Furthermore, the decrease in comment length for the NR group after the AI support was removed is also noteworthy. While longer comments do not necessarily guarantee higher quality, previous research has found a strong association between feedback quality and comment length Zong et al. (2021). Therefore, the decline in comment length for the NR group after the removal of AI support may indicate a decrease in the overall quality of their feedback. The current results reaffirm the findings from a prior study (Darvishi, Khosravi, Abdi, Sadiq, & Gašević, 2022) where students divided into AI-supported (AI) and non-AI-supported (No AI) groups exhibited notable differences. In that study, the AI group exhibited a flag rate that suggested moderate engagement with the AI prompts, along with a corresponding revision rate. Similarly, the AI group produced substantially longer comments than the No AI group, and their comments received higher likability for helpfulness. While decision and confidence ratings remained unaffected, the qualitative analysis of comments indicated that AI-supported feedback outperformed non-AI-supported feedback in terms of alignment with rubrics, specificity, explicit improvement suggestions, and constructive language. These findings demonstrate that the AI prompts played a significant role in maintaining the quality of students' feedback, and their influence declined after the support was removed. This suggests that AI can be an effective tool for scaffolding and automating learning activities. However, students tend to rely on AI assistance rather than learn from it, and their agency in providing high-quality feedback may be influenced by the presence or absence of AI prompts. This result is consistent with the literature on the idea of "distributed metacognition" – shared metacognition between the learner and the computer – and suggests that technology can increase learners' metacognitive resources (Kirsh, 2005). However, empirical research is limited in determining whether it improves metacognitive knowledge and independent self-regulation beyond technology interaction (Broadbent, Panadero, Lodge, & de Barba, 2020).

In response to RQ2, Table 3 shows evidence of the effectiveness of self-monitoring in maintaining student performance. Specifically, while the flag rate was still higher in the SR group compared to the AI group, it was lower than that in the NR group, with a smaller effect size. This suggests that the self-monitoring checklists helped students avoid making mistakes and provided them with a way to self-regulate their learning. The similarity and relatedness scores for the SR group showed negligible improvements compared to the NR group. However, the fact that there was no significant difference in comments length between the SR and AI groups is a

Table 4
Comparison between the control group (AI) and the third experimental group (SAI), where AI-assistance prompts are accompanied by the self-monitoring checklist.

AI vs SAI	AI (N = 396)		SAI (N = 418)		95% CI for Mean Difference		t	Cohen's d	p_{tukey}
Measure	Mean	SD	Mean	SD	Lower	Upper			
Flag Rate	0.14	0.23	0.14	0.24	−0.05	0.05	−0.10	−0.01	1.000
Similarity Score	0.28	0.12	0.28	0.12	−0.03	0.03	0.001	7×10^{-5}	1.000
Relatedness Score	0.31	0.13	0.31	0.13	−0.02	0.03	0.50	0.04	0.959
Comment length	23.38	12.41	25.67	14.75	−4.80	0.23	−2.34	−0.16	0.090
Time	265.44	988.78	267.77	880.74	−151.49	146.83	−0.04	−0.003	1.000
Like Rate	0.28	0.26	0.29	0.29	−0.08	0.02	−1.55	−0.11	0.408

notable achievement and an indicator of the success of the self-monitoring checklist in maintaining one of the "good behaviours" of providing a quality review as previous research found that providing lengthy comments, regardless of their quality, can benefit the reviewer and contribute to learning and self-regulation Zhu and Carless (2018); Baars et al. (2020). These findings support the effectiveness of self-monitoring checklists in maintaining student performance and promoting self-regulation in learning environments. Self-monitoring checklists can be an effective tool for helping students monitor and evaluate their own learning progress, which can lead to improved academic achievement and motivation. The considerable increase in the rate of flagged reviews in the NR group, where the system only recorded the flags without providing prompts, indicates that AI prompts played a significant role in student agency and the quality of their work. It suggests that avoiding the AI prompts might be a factor in keeping student agency to provide feedback that adheres to the guidelines. While the first experiment (NR) group's findings confirm our initial prediction that students rely on AI assistance to maintain the quality of their work, the second experiment group's data imply that a simple nudge, such as a checklist, could be beneficial in the absence of such support. Evidence from this study aligns with the existing literature on self-regulated learning and instructional scaffolding and underscores the potential of self-monitoring tools as effective aids for students in learning environments (Azevedo et al., 2008; Gašević et al., 2015, 2017).

In response to RQ3, our results did not yield support that the combination of self-monitoring checklists and AI prompts would lead to significant improvement in student peer review behaviour. The lack of significant advantages observed in combining self-monitoring checklists with AI prompts can be attributed to the principle governing the interaction between different forms of assistance. When a more potent and robust form of support is present, the contribution of a comparatively weaker approach may be overshadowed or rendered less impactful. For instance, imagine managing personal appointments with the assistance of a digital calendar that automatically adds meetings and provides reminders, along with a handwritten diary for personal notes. In this scenario, the digital calendar, as a stronger and more efficient form of assistance, may overshadow the contribution of a handwritten diary in managing personal appointments. This phenomenon can also be explained through the theory of cognitive load, which suggests that individuals have limited cognitive resources that can be overwhelmed if the load exceeds their capacity (Paas, Van Gog, & Sweller, 2010). In our study, the personalised AI assistance imposed a higher cognitive load on students due to its detailed prompts and guidance. This occupation of cognitive resources might have reduced the students' capacity to effectively utilise the self-monitoring checklists. Consequently, the combination of self-monitoring checklists and AI prompts yielded insignificant advantages, as the increased cognitive load associated with AI assistance might have diminished the potential impact of the self-monitoring strategy, resulting in the observed lack of significant benefits when the two approaches were combined.

However, we did observe some interesting trends in this study. Notably, students in the self-monitoring and AI group (SAI) provided the longest comments among all four groups. This finding suggests that students in the SAI group might have exerted more effort in their reviews, leading to longer comments, which have been associated with enhanced learning and self-regulation (Baars et al., 2020; Cavalcanti et al., 2020; Osakwe et al., 2021; Zhu & Carless, 2018). Fig. 3f shows the like rate – the number of comments from a user that received a like/helpfulness divided by the total comments submitted by that user. There are some interesting observations in this plot. First, removing the AI prompts in the NR group increased the chance of receiving likes on peer feedback, which might be due to the focus of the AI group on satisfying the criteria of the AI prompts rather than being constructive. Interestingly, replacing the prompts with the checklist increased the mean of the like rate in the SR group. Finally, while the addition of the checklist to the prompts in the SAI group seems to have had no significant change with similar median and upper quartile values in the chance of receiving likes on the comments, it considerably increased the lower quartile value and the average chance of receiving like/helpfulness from peers. While our previous study (Darvishi, Khosravi, Abdi, Sadiq, & Gašević, 2022) also demonstrated the positive impact of complementing peer review with AI prompts, checklists and training on peer review quality and increasing the chance of receiving helpfulness upvotes from peers, this study suggests the influence of a self-monitoring checklist in reducing the impact of the machine on students writings and increasing the humanisation of their feedback. There is still a need for improvement in the areas of the proper control distribution between learners and AI, as well as the creation of interfaces that enable reciprocal engagement. To ensure responsible AI integration in education, promoting the augmentation viewpoint and creating conversation among stakeholders, including academics, educators, developers, and policymakers, is critical (Molenaar, 2022b). We can picture a future where intelligent learning technologies successfully support students' agency and encourage meaningful learning outcomes by fusing the strengths of humans and AI and taking into account pedagogical considerations.

5.1. Limitations

There are several limitations to this study that should be considered when interpreting the results. First, the experiment was limited to the first eight weeks of the semester. It would be interesting to investigate the impacts of AI assistance on student agency in a longer-term context to see if the findings of this study hold over a longer period of time. Second, the no-support group (NR), which was compared to the AI group, had received AI assistance for the first four weeks, but it would be interesting to compare the quality of feedback between a group that never received AI assistance and a group similar to the NR group whose support has been removed after a limited time. This would provide a more fair evaluation of the effectiveness of AI prompts after they have been removed. Third, the study relied on quantitative measures of student performance, but it would be useful to supplement these findings with qualitative data collected through interviews or open-ended questions with students and instructors about their opinions of the AI tool and support. This could provide additional insight into the impact of AI assistance on student agency. Furthermore, the lack of significant improvement differences between the hybrid human-AI approach (SAI group) and the AI-only approach might be attributed to a potential design issue within the methodology. The AI assistance and self-checklist may not have been adequately tailored to accommodate the unique requirements of the hybrid approach, emphasising the need for future research to explore the design of

hybrid human-AI systems and their potential for enhancing student agency. Additionally, this study did not thoroughly investigate the potential challenges that AI-generated feedback can pose, such as the emergence of biases based on data and the introduction of new conflicts into the learning environment. It is essential to recognise that these challenges may have far-reaching consequences for both educators and learners when integrating AI into educational settings, highlighting the importance of future research to better understand and address these implications for AI in education. Also, while Appendix B reveals similar trends in student behaviour across different courses, mirroring the total dataset analysis, it is essential to recognise that variations in course content, instructional methods, student motivation levels, and individual instructor preferences could introduce additional complexities and nuances in the study's outcomes, representing an inherent limitation. Finally, it would be interesting to replicate this study with other learning tools and to provide AI support in tasks related to learning other than peer feedback. This could help to determine the extent of generalisability of the findings and to explore the impact of AI assistance on student performance in different contexts.

6. Conclusion

The increasing adoption of AI-powered educational technologies alongside contemporary instructional methodologies has shown potential for leveraging AI as a personal assistant that helps students by providing personalised reminders for completing tasks, automated real-time feedback for improving writing, or recommendations for when and what to study. However, our study in the context of providing AI assistance with writing feedback suggests that students tended to rely on AI assistance rather than actively learning from it. In our study, the reliance on AI became apparent when the assistance was removed, as students struggled to provide feedback of the same quality without the AI's guidance. Furthermore, our research explored the effects of replacing AI assistance with guidance and tips on providing constructive feedback, as well as the utilisation of self-monitoring checklists for students to evaluate the quality of their work independently. This approach proved to be effective, resulting in high-quality feedback when compared to the scenario of removing AI assistance entirely. However, it still demonstrated lesser effectiveness in comparison to relying solely on AI assistance. In addition, our investigation into hybrid human-AI approaches, which combine AI assistance with self-regulated strategies, revealed an interesting finding: supplementing AI assistance with self-regulated strategies did not yield significant advantages over relying solely on AI assistance. We speculate that this can be attributed to the principle that in the presence of a stronger form of assistance, the contribution of a weaker approach may be overshadowed or rendered less impactful.

In light of these findings, it is crucial to consider the broader implications and challenges associated with relying on AI assistance in the realm of education. While AI-powered learning technologies present numerous advantages, their implementation should be approached with caution, taking into account pedagogical considerations and carefully weighing the potential benefits against the possible drawbacks. Striking a balance between AI assistance and fostering student agency is essential to ensure that students actively participate in their own learning journey and develop crucial skills for the future. As we navigate a world where AI is increasingly integrated into our lives, it is vital to explore the best practices and ethical considerations surrounding the use of AI in education to maximise its potential while empowering learners.

Funding

This research was supported partially by the Australian Government through the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) project number IC200100022.

CRediT authorship contribution statement

Ali Darvishi: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft. **Hassan Khosravi:** Data curation, Investigation, Methodology, Writing - original draft. **Shazia Sadiq:** Supervision, Writing - review & editing. **Dragan Gašević:** Conceptualization, Writing - review & editing. **George Siemens:** Writing - review & editing.

Data availability

The authors do not have permission to share data.

A First four weeks

In the initial four weeks of the experiment, all participants were provided with AI prompts during the peer review process. During this period, a total of 1625 students participated in the study, collectively providing 16,007 peer reviews and evaluating 4501 resources across the ten courses. Table 5 provides an overview of the students and the peer review quantities for each group within the initial four weeks of the study.

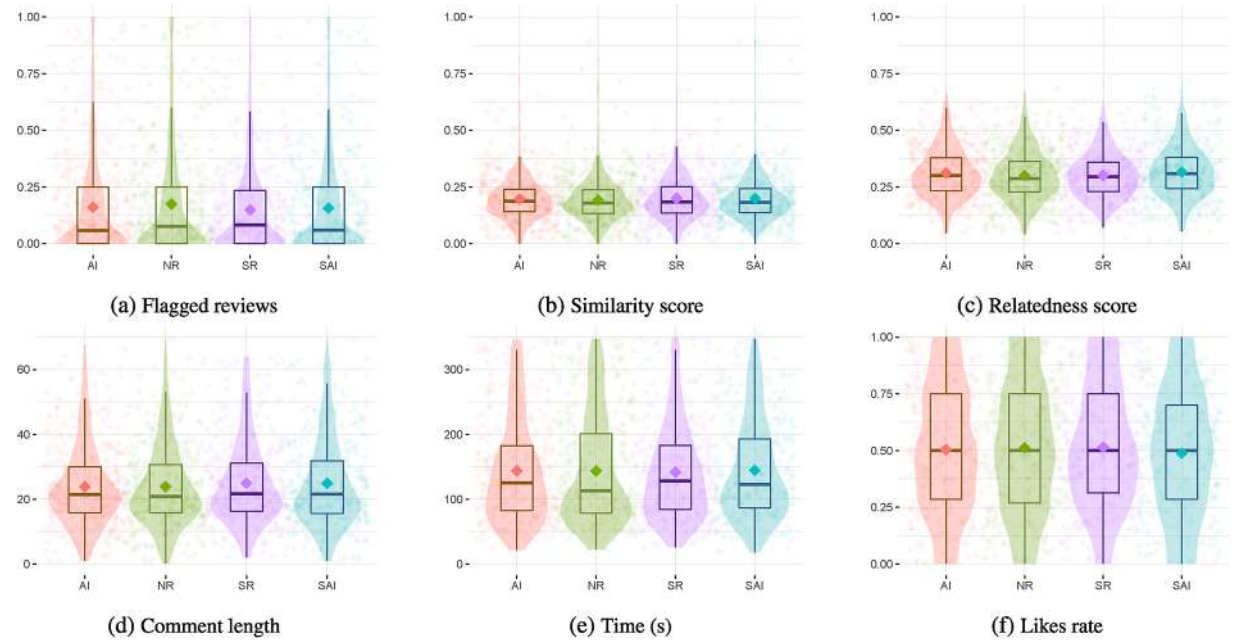


Fig. 4. Comparison of the control group (AI) and three experimental groups (NR, SR, and SAI) using six measures during the initial four weeks of AI prompt exposure. Boxplots represent data distribution with key statistics, while highlighted points indicate the mean. Violin plots provide an overview of data distribution using density curves.

Fig. 4 compares peer review behaviours among four groups (AI, NR, SR, and SAI) based on six measures detailed in Section 3.2.2 over the initial four weeks of exposure to AI prompts. Notably, statistical analysis revealed no significant differences in any of these measures between the groups during this period. Specifically, A one-way ANOVA revealed that there were no significant variations in the rate of flagged reviews ($F(3, 1621) = 0.866$, $p = 0.458$, $\eta^2 = 0.002$, Fig. 4a), average similarity scores with previous comments ($F(3, 1621) = 0.532$, $p = 0.660$, $\eta^2 = 0.001$, Fig. 4b), average relatedness scores of comments to reviewed resources ($F(3, 1621) = 2.04$, $p = 0.106$, $\eta^2 = 0.004$, Fig. 4c), average comment lengths ($F(3, 1621) = 1.92$, $p = 0.125$, $\eta^2 = 0.004$, Fig. 4d), average time spent on peer reviews ($F(3, 1621) = 0.586$, $p = 0.624$, $\eta^2 = 0.001$, Fig. 4e), or rates of received likes for helpfulness ($F(3, 1621) = 0.56$, $p = 0.641$, $\eta^2 = 0.001$, Fig. 4f). This lack of significant differences underscores the uniformity in peer review behaviour across the groups during the initial four weeks.

Table 5
Summary of the data collected during the initial four weeks, including the counts of students and peer reviews in each group, along with the total count of resources under review.

Peer review group	# Students	# peer reviews	# Resources
AI	396	3852	4501
NR	409	3926	
SR	402	3995	
SAI	418	4234	
Total	1625	16,007	

B Second four weeks results on course level

Table 6
Summary of the data collected in each course during the initial and second four weeks

Course Code	Description	School	# Students	First four weeks		Second four weeks	
				# Peer-reviews	# Resources	# Peer-reviews	# Resources
AGRC	Applied Mathematics & Statistics	Agriculture Food Sciences	243	3885	1121	1992	768
COMP	Artificial Intelligence	Info Tech & Elec Engineering	152	819	306	657	194
ECON	The Macroeconomy	Economics School	93	631	278	450	145
INFS	Introduction to Information Systems	Info Tech & Elec Engineering	177	1103	348	816	252
MEDI	Ethics and Professional Practice	Medical School	232	2537	941	1466	479
NEUR	The Brain and Behavioural Sciences	Psychology School	353	4024	540	3742	885
NUTR	Nutrition & Exercise	Human Movement & Nutrition Sciences	99	528	124	514	203
PHRM	Quality Use of Medicines	Pharmacy School	105	593	186	442	187
PSYC	Learning & Cognition	Psychology School	70	598	164	522	195
SLAT	Second Language Writing: T & R	Languages & Cultures School	101	1289	493	642	265
Total			1625	16,007	4501	11,243	3573

Table 6 provides an overview of various courses used in this study, including their codes, descriptions, associated schools, and student numbers. It also presents data on the number of peer reviews and resources used during the first and second four weeks of the courses. Fig. 5 presents an illustrative figure consisting of ten individual subplots, each representing the results for a specific course within our dataset. It is worth noting that while there may be variations in means and variances for each course, there is a consistent trend in student behaviour across all classes, echoing the results observed in the total dataset analysis (Fig. 3 in Section 4). In particular, it reveals that in all courses, the SR and NR groups consistently exhibit a higher rate of flagged reviews compared to the AI and SAI groups. This trend indicates that the presence of AI prompts and self-monitoring checklists consistently leads to a reduction in problematic reviews, regardless of the course. Furthermore, students in the SR and NR groups consistently produce comments that are more similar to their previous comments, whereas those in the AI and SAI groups maintain a lower level of similarity. Notably, students in the AI and SAI groups tend to provide longer comments across most courses, indicating that these interventions encourage more detailed feedback. Interestingly, in the INFS and MEDI courses, the SR group's performance in terms of comment length closely parallels that of the AI groups. However, it is crucial to highlight that these courses still follow the broader trend observed in the total dataset. The only area where some inconsistencies emerge is in the like rate, which measures the perceived helpfulness of reviews. While the overall data analysis revealed no significant differences in like rates between the groups, the variability in like rates across various courses and the occasional disparities in measures, such as comment length, emphasise further investigation to uncover the underlying factors contributing to these differences. Factors such as course content, variations in teaching methods, students' motivation levels, and individual instructor preferences may play pivotal roles in shaping these distinctions. Consequently, these findings serve as a catalyst for future research endeavours, encouraging a deeper exploration of these nuanced elements to enhance our understanding of AI role and peer review dynamics within diverse educational contexts.

Fig. 5. Comparing the control group (AI) with the three experimental groups (NR, SR, and SAI) using six measures for each course.

References

Abdi, S., Khosravi, H., Sadiq, S., & Darvishi, A. (2021, June). Open learner models for multi-activity educational systems. In *International conference on artificial intelligence in education* (pp. 11–17). Cham: Springer International Publishing.

Adie, L. E., Willis, J., & Van der Kleij, F. M. (2018). Diverse perspectives on student agency in classroom assessment. *Australian Educational Researcher, 45*, 1–12.

Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., et al. (2021). Explainable ai for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence, 4*, Article 723447.

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., et al. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative, Pharmacy, 19*, 1236–1242.

Arachchige, P. M., & Arosh, S. (2023). Large language models (llm) and chatgpt: A medical student perspective. *European Journal of Nuclear Medicine and Molecular Imaging, 50*, 2248–2249.

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition–implications for the design of computer-based scaffolds. *Instructional Science, 33*, 367–379.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research & Development, 56*, 45–72.
Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between student's effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review*, 1–24.
Bates, S. P., Galloway, R. K., Riise, J., & Homer, D. (2014). Assessing the quality of a student-generated question repository. *Physical Review Special Topics - Physics Education Research, 10*, Article 020105.
Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., et al. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology, 54*, 1222–1245.
Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies, 10*, 405–418.
du Boulay, B., Mitrović, A., & Yacef, K. (2023). *Handbook of artificial intelligence in education*. Edward Elgar Publishing.
Bowman, S. R. (2023). *Eight things to know about large language models*. arXiv preprint arXiv:2304.00612.
Bozkurt, A. (2023). Generative artificial intelligence (ai) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education, 18*.
Broadbent, J., Panadero, E., Lodge, J. M., & de Barba, P. (2020). Technologies to enhance self-regulated learning in online and computer-mediated learning environments. *Handbook of Research in Educational Communications and Technology: Learning By Design*, 37–52.
Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1–13.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., et al. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, Article 100027.
Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., et al. (2020). How good is my feedback? A content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 428–437).
Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*, 616–630.
Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 1*, Article 100002.
Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). *Electra: Pre-training text encoders as discriminators rather than generators*. arXiv preprint arXiv:2003.10555.
Code, J. (2020). Agency for learning: Intention, motivation, self-efficacy and self-regulation. *Frontiers in Genetics, 5*, 19.
Cooper, G. (2023). Examining science education in chatgpt: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*, 444–452.
Council, E. (2006). Recommendation of the european parliament and the council of 18 december 2006 on key competencies for lifelong learning. *Brussels: Official Journal of the European Union, 30*, 2006.
Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.S., & Gasevic, D., et al. (2023). Can large language models provide feedback to students? a case study on chatgpt. 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, 2023, pp. 323-325, doi: 10.1109/ICALT58122.2023.00100.
Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review, 64*, 313–342.
Darvishi, A., Khosravi, H., Abdi, S., Sadiq, S., & Gašević, D. (2022, June). Incorporating training, self-monitoring and AI-assistance to improve peer feedback quality. In *Proceedings of the Ninth ACM Conference on Learning@ Scale* (pp. 35–47).
Darvishi, A., Khosravi, H., Rahimi, A., Sadiq, S., & Gašević, D. (2022). Assessing the Quality of Student-Generated Content at Scale: A Comparative Analysis of Peer-Review Models. *IEEE Transactions on Learning Technologies, 16*(1), 106–120.
Darvishi, A., Khosravi, H., & Sadiq, S. (2020, September). Utilising learnersourcing to inform design loop adaptivity. In *European conference on technology enhanced learning* (pp. 332–346). Cham: Springer International Publishing.
Darvishi, A., Khosravi, H., & Sadiq, S. (2021, June). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the eighth ACM conference on learning@ scale* (pp. 139–150).
Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology, 53*(4), 844–875.
Darvishi, A., Khosravi, H., Sadiq, S., & Weber, B. (2022). Neurophysiological measurements in higher education: A systematic literature review. *International Journal of Artificial Intelligence in Education, 32*(2), 413–453.
Dawson, S., Joksimovic, S., Mills, C., Gašević, D., & Siemens, G. (2023). *Advancing theory in the age of artificial intelligence*.
Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, Article 104094.
Denny, P., Luxton-Reilly, A., & Simon, B. (2009). Quality of student contributed questions using peerwise. In *, 95. Proceedings of the Eleventh Australasian Conference on Computing Education* (pp. 55–63).
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv: 1810.04805.
Donker, T. (2023). The dangers of using large language models for peer review. *The Lancet Infectious Diseases, 23, 781*.
Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology, 103*, 962–1023.
Fyfe, P. (2022). *How to cheat on your final paper: Assigning ai for student writing* (Vols. 1–11). AI & SOCIETY.
Galloway, K. W., & Burns, S. (2015). Doing it for themselves: Students creating a high quality peer-learning environment. *Chemistry Education: Research and Practice, 16*, 82–92.
Gašević, D., Adesope, O., Joksimović, S., & Kovanović, V. (2015). Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The Internet and Higher Education, 24*, 53–65.
Gašević, D., Mirriahi, N., Dawson, S., & Joksimović, S. (2017). Effects of instructional conditions and experience on the adoption of a learning tool. *Computers in Human Behavior, 67*, 207–220.
Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. In *Computers and education, 100130doi*. Artificial Intelligence. https://doi.org/10.1016/j.caeai.2023.100130
Greene, J. A., & Azevedo, R. (2007). A theoretical review of winne and hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*, 334–372.
Heikkinen, S., Saqr, M., Malmberg, J., & Tedre, M. (2023). Supporting self-regulated learning with learning analytics interventions–a systematic literature review. *Education and Information Technologies, 28*, 3059–3088.
Henderson, M., Phillips, M., Ryan, T., Boud, D., Dawson, P., Molloy, E., et al. (2019). Conditions that enable effective feedback. *Higher Education Research and Development, 38*, 1401–1416.
Hilpert, J. C., Greene, J. A., & Bernacki, M. (2023). Leveraging complexity frameworks to refine theories of engagement: Advancing self-regulated learning in the age of artificial intelligence. *British Journal of Educational Technology, 54*, 1204–1221. https://doi.org/10.1111/bjet.13340.
HolonIQ. (2022). 2022 global education outlook. https://www.holoniq.com/notes/2022-global-education-outlook.
Hosseini, M., & Horbach, S. P. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research Integrity and Peer Review, 8*, 4.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 1*, Article 100001.
Inouye, K., Lee, S., & Oldac, Y. I. (2022). A systematic review of student agency in international higher education. *Higher Education, 1–21*.
Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology, 54*, 1057–1076. https://doi.org/10.1111/bjet.13325.
Jensen, L. X., Bearman, M., & Boud, D. (2021). *Understanding feedback in online learning–a critical review and metaphor analysis*. Computers & Education, Article 104271.
Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 1–20.
Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). *All-in-one: Multi-task learning bert models for evaluating peer assessments, EDM21*, 525–532. https://educationaldatamining.org/edm2021/.
Jin, S. H., Im, K., Yoo, M., Roll, I., & Seo, K. (2023). Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education, 20*, 1–21.
Kauffman, D. F., Zhao, R., & Yang, Y. S. (2011). Effects of online note taking formats and self-monitoring prompts on learning from online text: Using technology to enhance self-regulated learning. *Contemporary Educational Psychology, 36*, 313–322.
Khosravi, H., Demartini, G., Sadiq, S., & Gasevic, D. (2021). Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 32–42).
Kim, J., Nguyen, P. T., Weir, S., Guo, P. J., Miller, R. C., & Gajos, K. Z. (2014). Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 4017–4026).
Kirsh, D. (2005). Metacognition, distributed cognition and visual design. *Cognition, education, and communication technology*, 147–180.
Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., et al. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 15–24).
Krause, M., Garncarz, T., Song, J., Gerber, E. M., Bailey, B. P., & Dow, S. P. (2017). Critique style guide: Improving crowdsourced design feedback with a natural language model. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 4627–4639).
Kulkarni, C. E., Bernstein, M. S., & Klemmer, S. R. (2015). Peerstudio: Rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the second (2015) ACM conference on learning@ scale* (pp. 75–84).
Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *Albert: A lite bert for self-supervised learning of language representations*. arXiv preprint arXiv:1909.11942.
Li, R., Patel, T., & Du, X. (2023). *Prd: Peer rank and discussion improve large language model based evaluations*. arXiv preprint arXiv:2307.02762.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692.
Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*, 901.
Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P. C., et al. (2023). Chatgpt and large language models in academia: Opportunities and challenges. *BioData Mining, 16*, 20.
Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence, 5*, 333–334.
Misiejuk, K., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education, 175*, Article 104319.
Molenaar, I. (2022a). The concept of hybrid human-ai regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence, 3*, Article 100070.
Molenaar, I. (2022b). Towards hybrid human-ai learning technologies. *European Journal of Education, 57*, 632–645.
Moses, L., Rylak, D., Reader, T., Hertz, C., & Ogden, M. (2020). Educators' perspectives on supporting student agency. *Theory Into Practice, 59*, 213–222.
Negi, S., Asooja, K., Mehrotra, S., & Buitelaar, P. (2016). A study of suggestions in opinionated texts and their automatic detection. In *Proceedings of the fifth joint conference on lexical and computational semantics* (pp. 170–178).
Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.
Nieminen, J. H., Tai, J., Boud, D., & Henderson, M. (2022). Student agency in feedback: Beyond the individual. *Assessment & Evaluation in Higher Education, 47*, 95–108.
OECD, O.f. E. C.o. D. (2018). *The future of education and skills: Education 2030*. OECD Publishing.
Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gaševic, D., Cavalcanti, A. P., & Mello, R. F. (2021). Towards automated content analysis of feedback: A multi-language study. In *Proceedings of the 14th international conference on educational data mining*.
Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22*, 115–121.
Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561.
Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv preprint arXiv:1908.10084.
Rimbar, H. (2017). The influence of spell-checkers on students'ability to generate repairs of spelling errors. *Journal of Nusantara Studies (JONUS), 2*, 1–12.
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.
Schmitz, B., & Perels, F. (2011). Self-monitoring of self-regulation during math homework behaviour using standardized diaries. *Metacognition and Learning, 6*, 255–273.
Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International journal of educational technology in higher education, 18*, 1–23.
Shibani, A. (2019). *Augmenting pedagogic writing practice with contextualizable learning analytics*. Ph.D. thesis.
Stenalt, M. H., & Lassesen, B. (2022). Does student agency benefit student learning? A systematic review of higher education research. *Assessment & Evaluation in Higher Education, 47*, 653–669.
Suh, B. (2019). Can ai nudge us to make better choices?. URL: https://hbr.org/2019/05/can-ai-nudge-us-to-make-better-choices.
Tackett, S., Raymond, M., Desai, R., Haist, S. A., Morales, A., Gaglani, S., et al. (2018). Crowdsourcing for assessment items to support adaptive learning. *Medical Teacher, 40*, 838–841.
Tsai, Y. R. (2015). Applying the technology acceptance model (tam) to explore the effects of a course management system (cms)-assisted efl writing instruction. *Calico Journal, 32*, 153–171.
VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221.
Vaughn, M. (2020). What is student agency and why is it needed now more than ever? *Theory Into Practice, 59*, 109–118.
Walsh, J. L., Harris, B. H., Denny, P., & Smith, P. (2018). Formative student-authored question bank: Perceptions, question quality and association with summative performance. *Postgraduate Medical Journal, 94*, 97–103.
Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education, 191*, Article 104644.
Watkins, R. (2023). Guidance for researchers and peer-reviewers on the ethical use of large language models (llms) in scientific research workflows. *AI and Ethics*, 1–6.
Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist, 41*, 5–17.
Winne, P. H. (2013). Learning strategies, study skills, and self-regulated learning in postsecondary education. In *Higher education: Handbook of theory and research* (pp. 377–403). Springer.
WorldBank. (2021). Tertiary education overview. URL: https://www.worldbank.org/en/topic/tertiaryeducation.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv preprint arXiv:1609.08144.

Xia, Q., Chiu, T. K., & Chai, C. S. (2023). The moderating effects of gender and need satisfaction on self-regulated learning through artificial intelligence (ai). *Education and Information Technologies, 28*, 8691–8713.

Xiong, W., Litmaan, D., & Schunn, C. (2012). Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research, 4*, 155–176.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., ... & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *arXiv preprint arXiv:2303.13379*.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*, 1–27.

Zhu, M., & Bonk, C. J. (2019). Designing moocs to facilitate participant self-monitoring for self-directed learning. *Online Learning, 23*, 106–134.

Zhu, Q., & Carless, D. (2018). Dialogue within peer feedback processes: Clarification and negotiation of meaning. *Higher Education Research and Development, 37*, 883–897.

Zimmerman, B. J., & Paulsen, A. S. (1995). Self-monitoring during collegiate studying: An invaluable tool for academic self-regulation. *New Directions for Teaching and Learning, 13–27*, 1995.

Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior*, Article 106924.