Contemporary Educational Technology, 2024, **16**(4), ep541
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes

Sumie Tsz Sum Chan [1]
0009-0002-1352-1916

Noble Po Kan Lo [2,3*]
0000-0001-7636-6146

Alan Man Him Wong [1]
0009-0004-8988-2405

[1] English Language Teaching Unit, The Chinese University of Hong Kong, Hong Kong, CHINA
[2] Department of Educational Research, Lancaster University, Lancaster, UNITED KINGDOM
[3] Division of Languages and Communication, College of Professional and Continuing Education, The Hong Kong Polytechnic University, Hong Kong, CHINA
[*] Corresponding author: noble.lo@cpce-polyu.edu.hk

Citation: Chan, S. T. S., Lo, N. P. K., & Wong, A. M. H. (2024). Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes. *Contemporary Educational Technology, 16*(4), ep541. https://doi.org/10.30935/cedtech/15607

ARTICLE INFO

Received: 16 Jul 2024
Accepted: 12 Oct 2024

ABSTRACT

This paper investigates the effects of large language model (LLM) based feedback on the essay writing proficiency of university students in Hong Kong. It focuses on exploring the potential improvements that generative artificial intelligence (AI) can bring to student essay revisions, its effect on student engagement with writing tasks, and the emotions students experience while undergoing the process of revising written work. Utilizing a randomized controlled trial, it draws comparisons between the experiences and performance of 918 language students at a Hong Kong university, some of whom received generated feedback (GPT-3.5-turbo LLM) and some of whom did not. The impact of AI-generated feedback is assessed not only through quantifiable metrics, entailing statistical analysis of the impact of AI feedback on essay grading, but also through subjective indices, student surveys that captured motivational levels and emotional states, as well as thematic analysis of interviews with participating students. The incorporation of AI-generated feedback into the revision process demonstrated significant improvements in the caliber of students' essays. The quantitative data suggests notable effect sizes of statistical significance, while qualitative feedback from students highlights increases in engagement and motivation as well as a mixed emotional experience during revision among those who received AI feedback.

Keywords: LLMs, feedback, student engagement, student motivation, generative AI

INTRODUCTION

This paper examines the use of feedback generated by large language models (LLMs) on university students' writing proficiency. It focuses on examining the potential improvements in essay revisions across 918 first-year language students at a university in Hong Kong participating in a randomized controlled trial. Participants are assigned an argumentative essay task and then undergo revision processes, where some students receive feedback generated by artificial intelligence (AI) and others do not. This is undertaken to examine the potential for LLMs to foster writing skills in higher education, investigating in particular the potential for AI to provide useful feedback for students on their writing products. This introductory section

sets out the background to this study and discusses the rationale, aims, and research questions that it seeks to respond to.

Background

Artificial intelligence in education (AIEd) is a burgeoning sector within educational technology, offering potential benefits for large-scale teaching environments and providing real-time, personalized feedback to students (Gao et al., 2024). Despite AI's integration into applications over the past 30 years, ongoing research is essential to support large-scale teaching and intelligent assistance (Zawacki-Richter et al., 2019). Natural language processing (NLP), a subset of AI, has seen significant advancements in text processing, particularly with the development of transformer-based models like those used in self-attention mechanisms for NLP. The advent of powerful LLMs such as ChatGPT suggests a promising future for these technologies in education (Kasneci et al., 2023). The technical capabilities of automatic assessment systems have improved, and numerous studies highlight the potential of AIEd. Major advancements in AIEd can be categorized into four key areas (Zawacki-Richter et al., 2019): decision-making tools, intelligent tutoring systems, adaptive systems, and assessment and evaluation tools.

Decision-making tools aid in profiling and predicting admissions decisions, course scheduling, drop-out and retention rates, student modelling, and academic performance (Alvero et al., 2020; Chen et al., 2020; Langley, 2019). Intelligent tutoring systems are designed to teach course content, interact with students, curate learning materials, facilitate collaboration, and support teachers (Feng & Law, 2021; Hwang et al., 2020). Adaptive systems offer scaffolding and content personalization, help teachers understand student learning, use academic data to monitor and guide students, and represent knowledge through concept maps (Chen & Bai, 2010; Kabudi et al., 2021). Assessment and evaluation tools are used for automated grading, providing feedback, evaluating student understanding and engagement, ensuring academic integrity, and assessing teaching effectiveness (Huang et al., 2023; Luckin, 2017). These areas mark significant progress in AIEd.

This research is premised on the potential utility of feedback in improving student outcomes in written work. Extensive research supports the role of feedback in contributing to improved learning for students (Graham et al., 2015), and studies have also demonstrated that feedback that informs the revision process can improve grades on written work (Gnepp et al., 2020). However, providing feedback is a time-consuming process for teachers and marking is often cited as a source of teacher workload and stress (Hahn et al., 2021), with low-quality feedback also being a common complaint among university students (Madigan & Kim, 2021). This implies the need to improve feedback quality whilst reducing the workload upon teachers. In this regard, automated feedback presents a promising avenue to accomplishing this dual objective (Gao et al., 2024).

In terms of the potential of automated feedback programs to increase feedback consistency and reduce teachers' workload, some studies have already investigated the potential of automated writing evaluation (AWE) to reduce marking demands upon teachers (Crossley et al., 2022). Furthermore, a growing body of research investigates the potential for automated feedback to be applied by computer programs (Fleckenstein et al., 2023). However, previous attempts at developing such system have often focused on task-specific programs that are naturally limited in terms of their application for teachers on courses where teachers may pose a number of tasks (e.g., offering a choice of essay questions), or where assessment is based on broad criteria (e.g., opinion or reflection based writing tasks) (Ramesh & Sanampudi, 2022).

Rationale

The rapid improvement in AI technologies across the early 2020s have signaled the prospect for utilizing generative AI based on LLMs to both assess written work and to provide written feedback aimed at improving students' writing. As the literature review below reveals, research into the use of generative AI to provide feedback on students' work is growing, but the diverse nature of assessment across various educational systems and cultures means that there are often issues with the generalizability or transferability of findings. Through completing work in areas where there are current gaps in knowledge, the utility of generative-LLM AI can be evaluated with respect to its applicability in contemporary university level language education in Hong Kong.

Aims

This study aims at closing the research gap identified with respect to language education in Hong Kong. It is hoped that it is able to

(1) indicate the potential utility of AI in providing feedback on the written work of university level language students within this context and

(2) contribute to the broader literature on how AI is reshaping pedagogical and assessment practices.

It thus aims at contributing both to language education within universities in Hong Kong as well as carrying the discussion forward regarding the empirical basis supporting the use of AI in providing feedback on written work more broadly.

Research Questions

Meeting these aims requires designing a study that is suitably tailored to closing the gap in knowledge. In this vein, the study's research questions are, as follows:

1. To what extent can LLM-based generative AI provide feedback on written products that improves students' quality of work?
2. What are the experiences of students when receiving feedback from LLM-based generative AI, particularly in terms of their motivational, emotional, and attitudinal states?

This study responds to these research questions using a mixed-method design, utilizing both quantitative and qualitative methods of data collection and analysis. The details of how this is designed and the reasoning behind the specific design of the experiment and research instruments are given in the methodology section below.

Structure

The remainder of this study is structured in the following manner. First, a review of the literature is presented, discussing the research already carried out relevant to the research topic and identifying a gap in the literature. Following this, the study's methodology details the research methods and the rationale behind their design, as well as considering relevant ethical concerns. The results section puts forward separately the results to the quantitative analysis of participant in the experimental study, whilst the thematic findings of the qualitative analysis are also presented. These are discussed further in the section that follows before the study's findings are summarized and its contributions and limitations considered in the concluding section to the paper.

LITERATURE REVIEW

The extant body of literature on using LLM-based generative AI for feedback suggests there is potential utility for the application of AI to this end. Until relatively recently, studies largely focused on AWEs and their capacity to evaluate student work, though it has also been noted that its capacity for providing individualized feedback was limited (Mertens et al., 2022). The research findings regarding the effectiveness of providing feedback through LLM-based generative AI in improving student outcomes on revised work were mixed, with limitations observed in the feedback's applicability and specificity to the tasks that the AWE system is designed to evaluate (Fleckenstein et al., 2023).

By way of comparison, LLMS such as GPT have since emerged as a means to provide more tailored feedback to writing products (Yang et al., 2023). LLMs are trained on significant amounts of textual data, allowing them to generate natural language that mimics human feedback (Bowman, 2023). They are also capable of providing feedback on different types of work based on task inputs, learning objectives, and scoring systems, requiring relatively less coding time compared to AWEs (Bressane et al., 2024). There is therefore significant potential with respect to the application of LLMs such as GPT in providing automated feedback (Wardat et al., 2023).

However, as Chang et al. (2024) note, there is a lack of studies providing empirical support for the efficacy of the feedback generated by LLMs. Some have expressed concerns that AI-generated feedback might not be

accurate given that generative AI such as GPT often makes factual errors when completing generative tasks (Lee et al., 2024a). Likewise, some have observed that AI feedback in the hands of non-experts may not be as effective as feedback under research conditions given that student prompts might not be sufficiently detailed (Knoth et al., 2024). On the other hand, LLMs typically perform better at creative tasks, which may include providing feedback (Chia et al., 2023). Chang et al. (2024) also note that LLMs are capable of providing feedback without the use of reference texts and that they exhibit more potential for feedback than AWE.

This potential for LLMs to provide useful feedback is supported by some empirical studies. In comparing LLM and instructor feedback on written reports produced by university students, one study found that the AI-generated feedback was both coherent and broadly cohered with instructor feedback in terms of positive or negative assessments of the work (Dai et al., 2023). Other studies that use student or instructor evaluation of LLM-generated feedback report positive assessments of the technology and its utility on behalf of human participants (Jacobsen & Weber, 2023; Steiss et al., 2024). This may be balanced against some studies on the perspectives of English language teachers who express concerns about linguistic fidelity, overreliance on AI, and the suppression of student creativity (Al-Khreseh, 2024). However, there are limited empirical studies into measuring the impact of generative AI on student outcomes.

The few studies that have been carried out report promising results with respect to the utility of generative AI to provide helpful feedback on student work. One study on GPT feedback found that students who used AI to research their work (including using it for feedback) demonstrated better critical, reflective, and creative thinking skills than students who used traditional means of research and feedback (Essel et al., 2024). In a study by Meyer et al. (2024), 459 upper secondary EFL students were divided into two groups: one group received LLM-generated feedback and the other did not. The findings revealed that the written work written work evaluated by the AWE system showed greater improvement in the group that received feedback and revised their work, compared to the group that received non-AI feedback, suggesting the possibility of producing similar findings among English as a second language (ESL) students at the university level.

Beyond direct learning outcomes in terms of the scoring of written work, there are also other areas where AI feedback may be compared against instructor feedback. For instance, studies have shown that students' beliefs about the value of completing certain ESL writing tasks is linked to their motivation to complete such tasks (Eccles & Wigfield, 2020). Motivation is likewise related to positive student emotions, which have been found to be vital to the process of writing (Schrader & Kalyuga, 2020) and can be fostered by instructor feedback (Lipnevich et al., 2021). In this regard, students need to perceive feedback as effective in order for it to have a positive effect on emotions, motivations and engagement with specific tasks (Pandero & Lipnevich, 2022).

Fortunately, some preliminary evidence suggests that LLM-based feedback can indeed have a positive effect on student emotions. For instance, one study by Li and Xing (2021) demonstrated that LLMs could provide effective emotional support for students. Another study found that interaction with generative AI elicited positive perceptions and high levels of engagement (Aslan et al., 2024), though this was carried out with younger students. One study on the use of GPT with EFL learners found that its reframing of tasks could help foster greater cultural awareness among students, with positive responses from participating students (Zheng & Stewart, 2024). Importantly, a study by Al Shloul et al. (2024), which investigated the potential for GPT to improve student performance through feedback, found that most students saw its feedback as valuable and found interaction engaging.

However, what is less known based on the above studies is whether LLMs can provide feedback in a way that is perceived as effective by students, foster positive emotions, and motivate students to engage in work. This is particularly important given that feedback cycles, revision and submission can prove emotionally draining and demotivating for some students (McGarrell & Verbeem, 2007). The study carried out by Meyer et al. (2024) found moderate increases in task motivation and positive emotions, indicating the potential for LLM feedback to have beneficial emotional responses for participants. However, there is arguably also a need for qualitative research into these relationships in order to understand what aspects of AI-generated feedback students respond to in a positive (or negative) manner.

This literature review has highlighted several gaps in the literature. First, there are few studies that attempt to demonstrate the efficacy of LLM-based generative AI on student outcomes, though those that have been

carried out report positive correlations between the technology and student learning outcomes following revision of work. This highlights the need for more work in this area to establish connections that focus on specific educational contexts and areas of learning. Additionally, early indications that LLM feedback might be used to bolster student emotions and motivation require more qualitative research in order to better understand the mechanisms behind these relationships. These gaps inform the design of this research, as set out below.

METHODOLOGY

This section sets out the methodology of this study, justifying the selection of the experimental design used within the research, and detailing the methods of data collection and analysis used within the study. Following this, a brief discussion of research ethics will be presented.

Sample

The study collects data from 918 students enrolled in the first year of an English-language course at a higher education institute in Hong Kong. The students were all enrolled in a course that utilized international English language testing system and administered a task as part of a foundational university writings skills course. The students participating in the task were all Hong Kong citizens and English was their second language, while students who did not meet these criteria were filtered out. Of the samples, 55 percent were female and 45 per cent male, with the control groups being as representative as possible of this ratio. In total, 342 students were within the feedback group and 576 students were in the control group.

Experimental Design

The experiment took place in a two-hour lesson held in a computer laboratory on campus. Participating students were asked to complete the following writing task under test conditions:

> Do you agree or disagree with the following statement? Children under five ought to be prohibited from using tablet computers or smartphones. Use specific reasons and examples to support your answer.

A researcher was present throughout to prevent plagiarism and ensure that no student used generative AI to complete the task. Students were given 30 minutes to complete this task and then emailed their responses to the researcher.

Those in the feedback group had their work submitted by the researcher to GPT 3.5. This was preceded by a prompt setting out the task instructions and learning objectives and requesting no more than 500 words of feedback. All students received an email asking them to revise and improve their papers, with students in the feedback group receiving their LLM-generated feedback, and those in the control group receiving no feedback. They were given 5 minutes to prepare (and to read their feedback) and then a further 20 minutes to make revisions, before resubmitting their work to the researcher.

Both the original and revised papers were marked by instructors on the course. Manual scoring of work on behalf of instructors was undertaken due to the limited accuracy of LLMs and specifically GPT in scoring written student work (Lee et al., 2024b; Misiejuk et al., 2024). Papers and their revised papers were marked by separate instructors and all papers were double-marked with an average of the two marks constituting their final score.

Quantitative Methods

At the end of the experiment, all participating students were asked to fill in a short questionnaire about their experience. The questionnaire focused largely on their emotions and experiences with relation to the process of making revisions, asking them to describe how positive their emotions were during revision, how motivated they were to complete revisions, and how engaged they were with the process of revision. Scalar responses were collected that could then be compared across the control groups and measured against their scores from the writing task.

Analysis of the questionnaires took place in IBM's Statistical Package for the Social Sciences (SPSS) 29.0. This allowed for variables to be defined (e.g., gender, scalar variables, etc. and then cases created from data entered into the program(Salcedo & McCormick, 2020). Tests such as Pearson's product-moment correlation coefficient were used to identify numerical relations between sets of data, whilst the student's t-test was used to compare two or more groups' scores across a numerical variable (McCormick, 2015). An alpha of 0.05 was used across the tests, whilst the tests themselves were applied to data pertaining to test scores, questionnaire results, etc.

Qualitative Methods

Following each experiment session, an interview with a participating student from the feedback group was arranged to discuss their experience of LLM-generated feedback. In total, 16 of these interviews were successfully completed, lasting around an hour each. Interviews were selected because of their capacity to generate substantial information about individual perspectives as compared with questionnaires (Peters & Halcomb, 2015). Interviews were carried out by the researcher, who utilized a semi-structured approach to questioning, suitable for not only following the questions but also allowing the researcher to prompt the students for more details about areas of interest (Magaldi & Berler, 2020). The interviews were recorded on the researcher's tablet computer using digital audio recording software and then transcribed automatically using digital transcription software, before being manually corrected for any transcription errors.

The interview data was then subjected to thematic analysis. Thematic analysis serves as a means for identifying the themes raised by interviewees throughout the research process (Attride-Stirling, 2001). It focuses on identifying, analyzing and reporting patterns across data, describing and interpreting the themes prevalent across a dataset (Braun & Clarke, 2006). An approach to coding the data is required in order to complete this process (Guthrie, 2010). In this case, Leximancer was selected as a means for conducting thematic analysis of the interview data. Leximancer uses algorithms to extract semantic and relational data from the dataset and aggregate them into themes (Smith & Humphreys, 2006). These are represented in 'heat maps' that visually illustrate these themes, as well as other forums of output, such as ranked and co-occurring concepts (Smith & Humphreys, 2006). The unsupervised approach to coding and analysis was used in order to allow for the researcher to follow an inductive approach to analyzing interviewee responses to interview questions.

Ethical Considerations

Ethical considerations were considered when designing this research. For one, the British Educational Research Association's ethical guidelines for educational research (British Educational Research Association, 2018) were consulted when designing the study. Following its guidance, all participating students took part in the study voluntarily and were informed fully about their rights to withdraw from the study at any time. Their data was also anonymized at the point of marking and transcription, being attached only to a codename (e.g., student 1, 2, 3, etc.). Finally, the researcher considers the relative positions of power with respect to their relation to the students throughout the study, reflecting the need for positionality when undertaking primary qualitative research (Holmes, 2020).

RESULTS

This section presents the findings of the mixed-methods approach taken to study within this paper. It presents first the results of the quantitative analysis of data and then the findings of the qualitative analysis of interviews. These findings are discussed in more depth in relation to the study's aims and research questions in the discussion section that follows.

Quantitative Analysis

In exploring the results for the tests and questionnaires across the feedback and control group, it is clear from **Table 1** that revised scores for the feedback group were higher than that for the control group, with the feedback group receiving roughly 3.113 extract marks on their revised paper than the controlled group.

Table 1. Average scores for task and questionnaires across and between feedback and control groups

Group	Task score	Revised score	Difference	Emotion	Motivation	Engagement
Feedback	56.402060	63.989690	7.587629	4.556701	5.051546	4.969072
Control	56.680410	61.154640	4.474227	3.762887	3.350515	3.989691
Difference	–0.278350	2.835052	3.113402	0.793814	1.701031	0.979381

Table 2. t-test results

	Mean difference	df	t-value	p-value
Improvements	3.113	192	2.947	0.00360
Emotion	0.794	192	1.884	0.06110
Motivation	1.701	192	4.193	0.00004
Engagement	0.979	192	2.089	0.03460

Additionally, the feedback group self-reported higher levels of emotion, motivation and engagement, with the motivation score being 1.7 points higher on a scalar score of 1 to 10.

Applying the t-test to these scores was used to ensure that there were sufficient statistical differences between the two groups (**Table 2**). In terms of the differences between the two scores, the feedback group saw a mean improvement of 7.588 (standard deviation [SD] = 7.477) and the control group a mean improvement of 4.474 (SD = 7.157), resulting in a p value of 0.003604, indicating a high likelihood that the null hypothesis can be rejected. In comparing the positive emotion scores, the mean score for the feedback group was 4.557 (SD = 3.07) and a lower 3.763 for the control group (SD = 2.761). The difference here was the lowest of all self-reported scores at 0.794 and was not found to be statistically significant (p = 0.0611). The differences for motivation and engagement scores were larger, however, with a 1.701 score difference in motivation proving statistically significant (FG SD = 3.108, CG SD = 2.479, p = 0.00004) and a 0.979 difference in engagement score also proving statistically significant (CG SD = 3.411, CG SD = 2.951, p = 0.0346). There were thus statistically significant improvements in the scores of the revised papers and in experiences of motivation and engagement when comparing the feedback and control groups.

The size of the effects between the groups may be explored beyond the differences between the means above by establishing coefficient scores. To explore the relationship between receiving feedback versus not receiving feedback and the difference between the original and revised paper scores, a point biserial correlation was calculated, generating a coefficient of 0.208, indicating a weak positive correlation between received AI-generated feedback and increased improvements in test scores. Comparing the groups who received and did not receive feedback, there was a very weak positive correlation between receiving feedback and positive emotion (0.135) and likewise with engagement (0.152), although there was a larger weak-to-moderate correlation with motivation (0.29). There were thus positive correlations between feedback and all measures, though statistical significance needs to be considered.

Attempting to establish the mechanisms at work in the different groups requires understanding how correlated variables such as emotion, motivation and engagement are with the improvements in scores. In order to achieve this, Pearson's correlation coefficient was calculated to compare effect and significance between these variables and the differences between the two scores. Taking the group as a whole, the effect of emotional score upon score differential between the two papers was 0.543, indicating a moderate effect that was statistically significant ($p = 1.83188 \times 10^{-15}$). The effect sizes of motivation (r = 0.882) and engagement (r = 0.883) were very strong, again, with high statistical significance ($p = 1.24235 \times 10^{-64}$ & $p = 5.90325 \times 10^{-65}$). As **Figure 1** shows, the effect of all three experience scores on revised test score improvements increased exponentially, implying that positive affective experiences of feedback and revision became increasingly valuable as they became more enjoyable. It may therefore be theorized that the strong relationships between motivation and engagement and revised scores–coupled with the weak effect of feedback on motivation and engagement–account for the larger relative increase in the scores received for the revised paper among the feedback group.

To confirm these findings, additional tests were conducted to measure the effect of engagement and motivation on test scores. Simple linear regression tests were performed for both the feedback and control groups (**Table 3**).

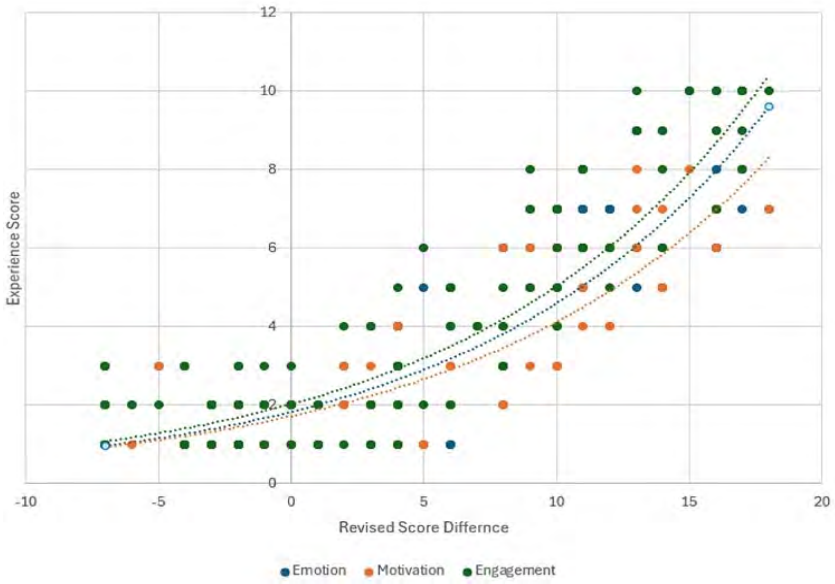

Figure 1. Self-reported scores for experiences (positive emotions, motivation, and engagement) cross-referenced with associated score improvement on revised task submissions (Source: Research project database)

Table 3. Regression analysis

Variable	Value			
	Motivation effect on score difference for control group	Engagement effect on score difference for control group	Motivation effect on score difference for feedback group	Engagement effect on score difference for feedback group
Multiple R	0.842	0.873	0.911	0.889
R-squared	70.1%	76.2%	83.0%	79.0%
Standard error	1.351	1.407	1.293	1.581
Significance F	3.388×10^{-27}	2.07×10^{-31}	2.19×10^{-38}	6.38×10^{-34}
Multiple R	0.842	0.873	0.911	0.889

In the case of the effect of engagement on score improvements, a strong correlation was observed, with an improvement of 0.842 and a fit of 70.1% for the control group, indicating a strong correlation between scores and considerable influence of motivation upon these scores. However, the correlation was even more substantial for the feedback group, where motivation likewise accounted for 10% more of their score differentials. Similar results were displayed when the effect of engagement was calculated with the Multiple R increasing slightly from 0.873 to 0.889 and the R-Square rising from 76.2% to 79%. However, the feedback group saw a bigger jump in the effect of motivation than engagement, suggesting that feedback enhances motivation far more than engagement.

Qualitative Analysis

Thematic analysis of interviews with sixteen students was carried out through *Leximancer,* using the program's in-built algorithm to code and organize themes. This method of analysis produces themes based on the frequency, proximity and semantic connections between terms used in the interview transcripts. A concept map (**Figure 2**) demonstrates some of the concepts associated with the interviews and their association with related concepts. The six main concepts identified through this process are: feedback, paper, having, task, better, and forward.

The prevalence of the concepts within the Venn diagram is listed more clearly in **Figure 3**, which signals their frequency across the interviews.

These thematic concepts identified in **Table 2** contain themselves other sub-concepts. Here, some related concepts are grouped together–for instance, 'writing', 'revision', 'felt', 'feel', and 'experience' are grouped under 'feedback', whereas 'suggestions', 'help', 'motivation', 'provided' and 'improved' are categorized under 'paper'.

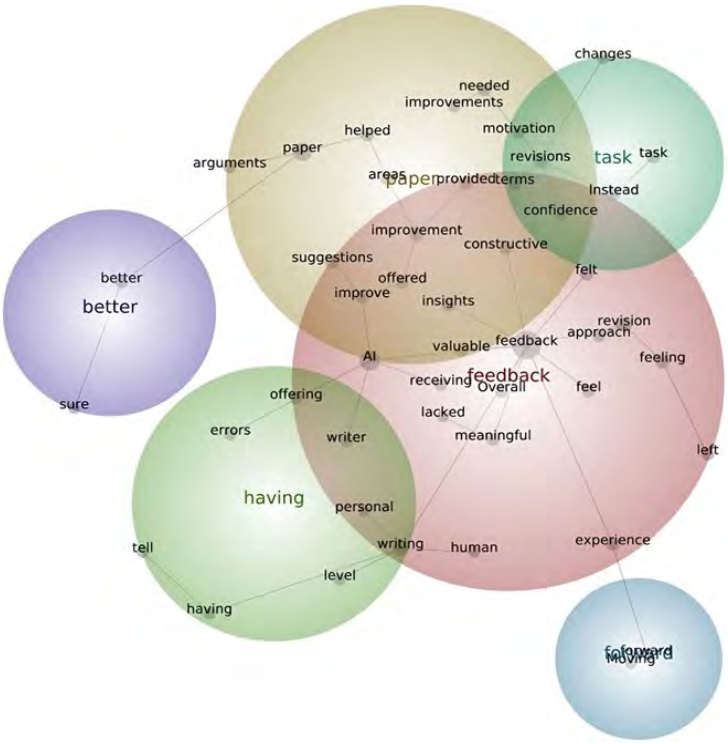

Figure 2. Concept map for analysis of interviews derived from *Leximancer* analysis (Source: Research project database)

Theme	Hits
feedback	208
paper	150
having	73
task	49
better	16
forward	10

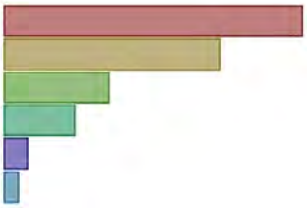

Figure 3. Major themes with frequency of conceptual frequency (Source: Research project database)

This indicates the relationship of concepts to each other in the responses, suggesting that the concepts associated with other concepts across the responses. This is developed upon below through looking at concept pathways and the excerpts attached to certain themes identified by the analysis. Thematic concepts identified in **Table 2** contain themselves other sub-concepts whose frequency is outlined in **Figure 4**.

With respect to the theme of 'feedback', the respondents utilized concepts such as AI, writing, revision, feeling, felt, insights, feel, experience, meaningful, valuable, receiving, approach, left, human, and lacked. Here, emotions were clearly attached to the experience of AI. Many of these were positive but some were negative. For example:

> Emotionally, receiving feedback from an AI left me feeling somewhat disengaged and uninspired. Unlike feedback from a human instructor or peer, which can be personalized and empathetic, the AI feedback felt impersonal and detached.

In other cases, feedback was more positive, with statements such as 'the feedback made me feel like I could achieve better than what I already had, so I sort of sat up and took notice when I got'. However, it was notable that discussion of feelings across the interviews were not wholly positive, with a mixed response to the experience of receiving AI feedback. Some compared it negatively to human feedback whilst only one assessed it positively in comparison to human feedback.

The connections between emotional concepts and other concepts indicate how the interviews saw emotional responses to AI linked to improved outcomes.

	Count	Relevance				
feedback	193	100%	provided	17	9%	
paper	77	40%	insights	17	9%	
writing	55	28%	improve	17	9%	
revisions	40	21%	constructive	17	9%	
revision	36	19%	errors	16	8%	
suggestions	29	15%	left	14	7%	
felt	26	13%	tell	14	7%	
helped	25	13%	offered	13	7%	
feeling	25	13%	meaningful	13	7%	
improvement	23	12%	improvements	13	7%	
task	23	12%	confidence	13	7%	
feel	23	12%	arguments	13	7%	
experience	23	12%	terms	12	6%	
having	23	12%	approach	12	6%	
areas	20	10%	personal	12	6%	
needed	20	10%	writer	12	6%	
motivation	19	10%	valuable	11	6%	

Figure 4. Ranked concepts derived from Leximancer analysis of interviews with participating students (Source: Research project database)

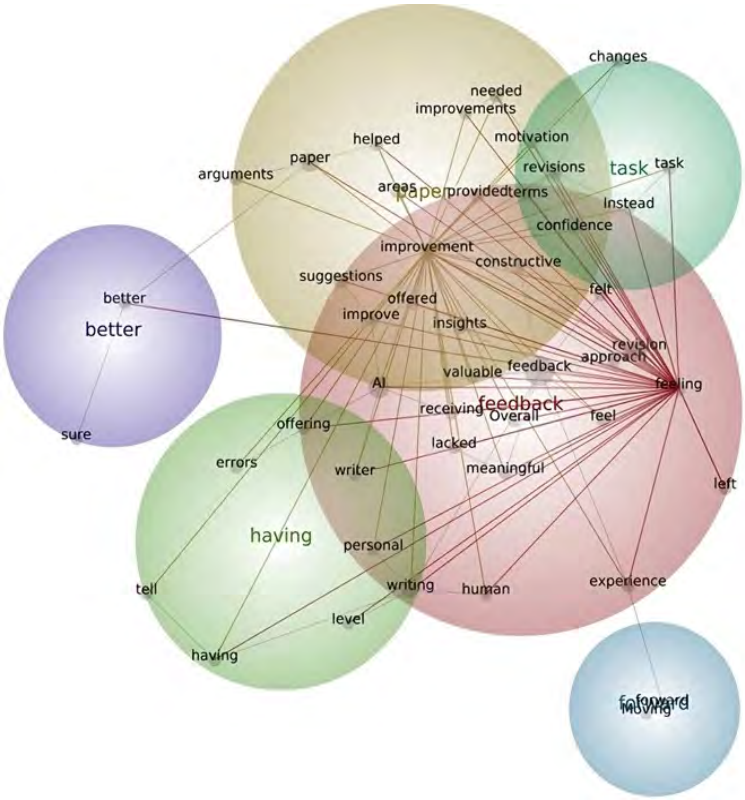

Figure 5. Pathways between 'feeling' and 'improvement' (Source: Research project database)

Figure 5 demonstrates how the words 'feeling' and 'improvement' were linked through the concepts, such as 'suggestions' and 'helped', which were among the most prevalent terms linking the two. This can be evinced from statements such as this: 'The suggestions left me feeling like I could have improved what I had submitted' and 'The AI bit really left me feeling like I knew where to go in terms of what I had to do next'. However, arguably these connections were not as strong as were connections between emotions and specific emotional statements in terms of concepts such as 'sad', 'delighted', and 'irritated'.

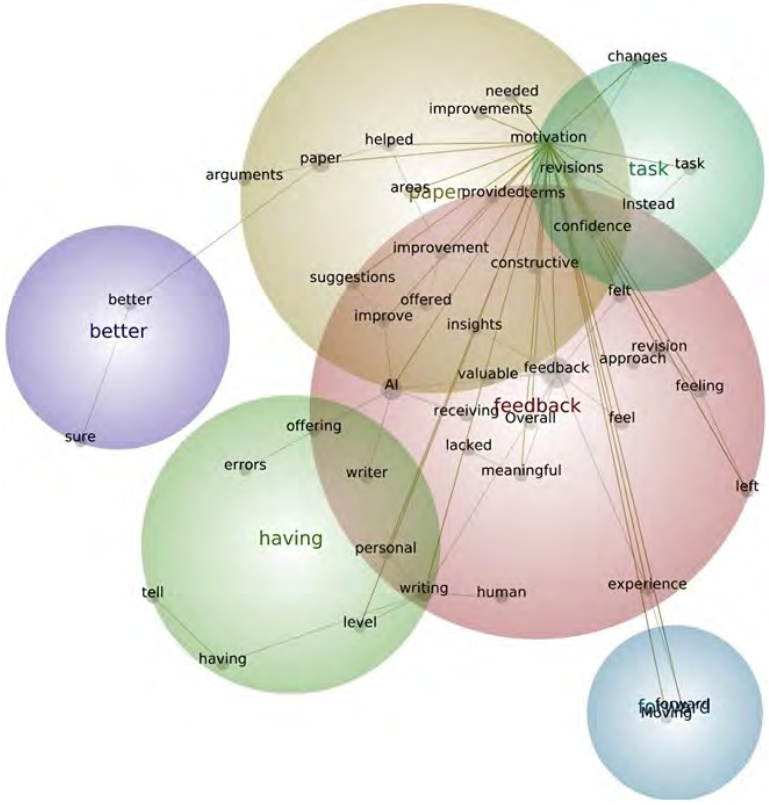

Figure 6. Pathways between 'motivation' and related concepts (Source: Research project database)

The participants' responses in terms of describing their feelings of motivation again demonstrate a mixture of responses. Some were lukewarm with respect to the effect of the feedback whereas others were enthusiastic: 'I guess it gave me some motivation to keeping going, but it wasn't like I was, you know, super excited about it or anything'; 'Yeah, in terms of motivation, the feedback really made it seem worthwhile. Like, before I was dreading going over the same paper, but now there was kind of a point to it–it gave me a new task'. 'Excitement' was mentioned by some in relation to motivation, indicating a potential connection between emotional states and motivation.

The concept pathway map for 'motivation' demonstrates its intersection with other concepts, with 'provided', 'revision/s' 'needed' 'improvement/s' 'help/ed' and 'change/s' ranking highly among the related concepts (**Figure 6**). In addition, links to other concepts such as '[moving] forward/s' and 'level' suggest that the perceived effect of motivation was to improve the level of revision the participants felt they were able to produce. Interestingly, despite the similarly high levels of effect for motivation and engagement cited in the statistical analysis above, there were few linkages between motivation and engagement in the interviews. Indeed, engagement as a concept was not identified by *Leximancer* as a common concept, though it may have been conflated with 'motivation' under its algorithmic analysis.

Other themes identified above such as 'having' likely result from the prompt of the questions to describe experiences, with many responses beginning with the phrase 'it was like having', such as 'it was like having a lecturer actually give me feedback' and 'once I could see the value in it, it was like having a second pair of eyes have a look at my paper'. The term 'paper' was a major concept identified by the analysis, often with connection to 'motivation', such as 'receiving feedback from the AI definitely made me feel more motivated as I could see what I actually needed to apply to improve my paper'.

Some described the feedback as a 'roadmap' to improving their papers, acknowledging that the feedback has contributed to their 'knowledge': 'Knowing exactly what I needed to address let me focus on those goals for the twenty minutes we had'. Indeed, the specificity of the feedback tailored to the individual was remarked upon as helpful in this regard, with one participant stating, 'I was honestly impressed with how well it had able to identify the points within my paper, so shock was probably an emotional reaction'.

The term 'task' was connected to others such as 'insights', 'improved', and 'confidence', indicating that the relevance of the feedback to the task was noted. Indeed, few participants contradicted the view that the feedback was sufficiently relevant to the task, signaling perhaps the utility of AI to responding to paper content without the need to be designed for a specific task as in the case of AWEs. This functionality of the LLM-based feedback does not appear to be disputed across the interviews, though the experience of receiving AI feedback–particularly with regards to emotional responses–does vary substantially.

Indeed, when examining the concept 'better', there are negative as well as positive concepts attached to it. For instance, one student stated that 'I feel I did definitely perform better with the feedback, but I still don't like getting instruction off an AI. It feels weird'. 'The word 'arguments' associated with 'better' in 15 per cent of cases suggested that the participants felt that it helped them improve their arguments in the research, though the absence of terms such as 'English' or 'language' suggests that the language students did not feel that it particularly improved the linguistic content of their work. As the task was general rather than language-based, however, perhaps this is the reason for this omission.

Summary

The findings identified across the quantitative and qualitative analyses revealed statistically significant correlations between receiving AI-generated feedback and increases in grade between original and revised papers. The effect of this is small but statistically significant, whereas correlations between receiving feedback and motivation and engagement demonstrate a slightly larger effect that is statistically significant. The quantitative analysis further indicates that positive emotions and especially motivation and engagement can have moderate-to-large effects on revision performance, whilst the interviews suggest that effects on motivation are the most pronounced, as the respondents associated it with improvements in their papers. Emotional responses were also mixed and may not necessarily be as correlated with motivations to complete revisions as may otherwise be causally assumed. These results and findings are discussed in more depth in the section that follows.

DISCUSSION

The above results demonstrate a number of trends that warrant discussion. Quantitative analysis indicates that the feedback group received higher revised scores and reported higher levels of emotion, motivation and engagement when compared with the control group. T-tests revealed that these differences were statistically significant, with the exception of the differences in emotional responses to feedback. The effect of feedback on positive emotions was also less pronounced according to the point biserial correlation analysis of the effects of feedback on emotion, as well as the Pearson's correlation coefficient examining the relationship between emotional and revision scores generally.

The interviews likewise revealed a mixed emotional response to receiving feedback, with participants reporting dissatisfaction with how receiving AI feedback left them feeling disengaged and uninspired. This contradicts the findings of similar research which indicate that LLM-based feedback can have a statistically significant positive effect on student affective experience and emotional well-being (Li & Xing, 2021). This may be because feedback by nature is critical, though some did identify a perception of AI feedback as not comparable with human feedback or as lacking in some regard in comparison. Previous studies have also found that feedback cycles can prove emotionally exhausting for students (McGarrell & Verbeem, 2007), and it is possible that this effect is more pronounced with LLM-based feedback or was in some way exacerbated by the design of the study.

Interestingly, this did not appear to be reflected in responses regarding motivation, which were broadly positive. Motivation was also linked by the interviewees to improvements made during revision, reflecting perhaps the strong correlations identified by the Pearson's correlation coefficient between motivation and revision scores and between engagement and revision scores. Prior research has indicated that motivation is influential with respect to measured outcomes during writing tasks (Schrader & Kalyuga, 2020), whilst other research has suggested that instructor feedback can successfully enhance student motivation (Lipnevich et al., 2021). The above findings seem to suggest that not only can LLM-based feedback enhance motivation significantly, but this also accounts for a substantial proportion of observed differences in test scores.

With respect to the effects of AI-generated feedback on these scores, there was also a statistical correlation between receiving AI feedback and the difference between the marks given to the original written product and those to the revised written product. Using a t-test to compare the differences in scores for the control and feedback groups, there was a statistically significant difference between the two groups, with a point bivariate correlation analysis revealing a weak positive effect of receiving feedback on outcomes. This goes some way towards closing the gap observed by Chang et al. (2024) regarding the sparse empirical evidence supporting the effect of LLM-based feedback on learning outcomes.

It can be inferred that LLM-based feedback has a positive effect on test scores through providing sufficiently targeted feedback. Even among students who reported negative impacts on emotion, there was a consensus in the interviews that the feedback was useful and targeted, implying that its mechanism occurs more through motivation and engagement than through shifts in emotional state. One possible explanation is that the relevant shift in mental states is in terms of attitudinal disposition towards the task, with the interviewees reporting that they felt the feedback gave the task purpose or meaning, as well as citing the usefulness of having specific, actionable feedback at their disposal.

However, there is a possibility that these findings are not transferrable into real-world scenarios. Motivation in the context of the task may have been improved by feedback as it provides a purpose for the revision rather than for the utility of its specific guidance. As the tasks were not creditable in terms of contributing to a course of study or qualification, revising the paper may have only appeared to have a point in light of new feedback and 'instructions' from AI. Comparing pre- and post-feedback emotional and attitudinal states is unfortunately not possible as there was no pre-feedback questionnaire, meaning that changes in motivational state before and after the intervention cannot be compared. The nature of the interview analysis also inhibits insights into these relationships given that the output of the algorithm pertains largely to the semantic content of responses rather than the specific views, attitudes and experiences of participants.

Another limitation to the study is with respect to what it says about domain-specific knowledge and written tasks within them. The task assigned to the English-language students was fairly generic and not related clearly to subject-specific skills. The respondents to the interviews naturally did not report how far the input from AI improved their writing skills and instead focused on its ability to present them with new arguments or content. This perhaps reflects concerns about the limitations of AI in terms of serving well to create content but not as well when it comes to its analytical function (Knoth et al., 2024). It may be, for example, that the content of the feedback received was not particularly helpful and that its association with relatively improved scores is attributable wholly to its effect on motivation and engagement. In other words, the design of the study does not allow for any evaluation of how accurate, relevant or helpful the actual guidance provided by the AI proved.

Nevertheless, the statistical correlations demonstrate that the group that received feedback from AI did indeed experience improvements in their scores relative to those of the control group. This suggests that whilst there is not yet sufficient information to develop a clear model for pattern of causal influence behind the relationship, there is indeed a positive relationship that warrants further research controlling for variables. This is reflected in the recommendations offered below.

The findings suggest that LLM-based generative AI has the potential to significantly improve the quality of students' written work. The quantitative analysis demonstrates that students who received AI-generated feedback performed better when revising their written papers as compared with the control group, scoring on average 3.113 marks higher. Statistical tests confirm that these improvements were statistically significant and that there was a weak effect ($r = 0.208$) of receiving AI feedback on subsequent performance relative to the control group. Motivation and engagement also had strong positive correlations with score improvements, whilst the feedback group enjoyed a weak improvement in both measures as compared with the control group. Emotional positivity was found to be correlated with scores to a moderate degree but had weaker or insignificant relationships with feedback across other measures.

Students' experiences with LLM-based AI feedback were mixed in terms of the emotional response, with interviews noting a mixed response whilst statistical analysis denied a significant link between feedback and emotional positivity. Relationships between feedback and motivation and engagement were statistically

significant but weak in strength, though both motivation and engagement were strongly correlated with general revision performance across both groups. It is therefore possible that AI feedback functioned through increasing student motivation and encouraging engagement during the task, though it is unclear whether this finding is transferrable to real-world scenarios. Likewise, it is possible that other latent variables play a role in the noted improvements in scores following the intervention.

CONCLUSION

The study indicates that AI feedback modestly enhances the revision of written work, whilst suggesting that slight improvements in motivation and engagement may partly account for this relationship. Future research could build on these findings through investigating how AI's personalization has an impact upon emotional states, given the variability of findings in this regard. This could perhaps consider pre-intervention attitudes towards AI and how they relate to subsequent effects on emotions, attitudes, and performance. Negative comparisons of the experience of AI feedback as compared with human feedback invite further research comparing the experience of both types of feedback, as well as towards assessing the relevance, accuracy and effect of AI feedback as compared with instructor feedback.

Whilst the study indicates the potential of AI to provide valid feedback under experimental conditions, longitudinal studies that measure AI feedback use in real-world educational scenarios would go a long way to establishing the transferability of these findings. Additionally, investigations into written products with respect to domain-specific assessments might establish where LLM-based AI is most effective in terms of subject domains and assessment types. In this regard, attempting to differentiate between the focus of AI feedback and its effects–e.g., on writing skills, content, language skills, etc.–can help better understand the mechanisms of its effect as well as identifying for whom AI feedback is best suited.

Author contributions: STSC, NPKL, AMHW: development and execution of the study; **STSC:** data collection, expertly managing data entry into spreadsheets and converting qualitative data into quantitative form, thereby solidifying the empirical foundation of the study; **NPKL:** initiated the project, conceptualized the study framework, analyzed the data, and prepared the initial manuscript draft; **AMHW:** contributed to the discussion sections, providing critical insights that enhanced the theoretical depth of the study. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: The authors declared that this study did not require an ethics committee approval since this study did not include any personal information. The authors further declared that the highest ethical practices were followed during the study.

Declaration of interest: The authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Al Shloul, T., Mazhar, T., Abbas, Q., Iqbal, M., Ghadi, Y. Y., Shahzad, T., Mallek, F., & Hamam, H. (2024). Role of activity-based learning and ChatGPT on students' performance in education. *Computers and Education: Artificial Intelligence, 6*, Article 100219. https://doi.org/10.1016/j.caeai.2024.100219

Al-Khreseh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence, 6*, Article 100218. https://doi.org/10.1016/j.caeai.2024.100218

Alvero, A. J., Arthurs, N., Antonio, A. L., Domingue, B. W., Gebre-Medhin, B., Giebel, S., & Stevens, M. L. (2020). AI and holistic review: Informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 200–206). ACM. https://doi.org/10.1145/3375627.3375871

Aslan, S., Durham, L. M., Alyuz, N., Okur, E., Sharma, S., Savur, C., & Nachman, L. (2024). Immersive multi-modal pedagogical conversational artificial intelligence for early childhood education: An exploratory case study in the wild. *Computers and Education: Artificial Intelligence, 6*, Article 100220. https://doi.org/10.1016/j.caeai.2024.100220

Attride-Stirling, J. (2001). Thematic networks: An analytical tool for qualitative research. *Commission for Health Improvement, 1*(3), 385–405. https://doi.org/10.1177/146879410100100307

Bowman, S. R. (2023). Eight things to know about large language models. *arXiv*. https://doi.org/10.48550/arXiv.2304.00612

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bressane, A., Zwirn, D., Essiptchouk, A., Saraiva, A. C. V., de Campos Carvalho, F. L., Formiga, J. K. S., de Castro Medeiros, L. C., & Negri, R. G. (2024). Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: A proposal using artificial intelligence. *Computers and Education: Artificial Intelligence, 6*, Article 100196. https://doi.org/10.1016/j.caeai.2023.100196

British Educational Research Association. (2018). *Ethical guidelines for educational research*. British Educational Research Association.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Wang, C., Wang, Y., Ye, W., Zhang, Y., Zhang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evolution of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3), Article 39. https://doi.org/10.1145/3641289

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Chen, S.-M., & Bai, S.-M. (2010). Using data mining techniques to automatically construct concept maps for adaptive learning systems. *Expert Systems with Applications, 37*(6), 4496–4503. https://doi.org/10.1016/j.eswa.2009.12.060

Chia, Y. K., Hong, P., Bing, L., & Pira, S. (2023). Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv*. https://doi.org/10.48550/arXiv.2306.04757

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Banner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse element (PERSUADE) corpus 1.0. *Assessing Writing, 54*, Article 100667. https://doi.org/10.1016/j.asw.2022.100667

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61*, Article 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence, 6*, Article 100198. https://doi.org/10.1016/j.caeai.2023.100198

Feng, S., & Law, N. (2021). Mapping artificial intelligence in education research: A network-based keyword analysis. *International Journal of Artificial Intelligence in Education, 31*, 277–303. https://doi.org/10.1007/s40593-021-00244-4

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*. https://doi.org/10.3389/frai.2023.1162454

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education. *Computers and Education: Artificial Intelligence, 6*, Article 100206. https://doi.org/10.1016/j.caeai.2024.100206

Gnepp, J., Klayman, J., Williamson, I. O., & Barlas, S. (2020). The future of feedback: Motivating performance improvement through future-focused feedback. *PLoS ONE, 15*(6), Article e0234444. https://doi.org/10.1371/journal.pone.0234444

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Guthrie, G. (2010). *Basic research methods: An entry to social science research*. SAGE. https://doi.org/10.4135/9788132105961

Hahn, M. G., Navarro, S. M. B., La Fuente Valentin, I., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access, 9*, 108190–108198. https://doi.org/10.1109/ACCESS.2021.3100890

Holmes, A. G. D. (2020). Researcher positionality–A consideration of its influence and place in qualitative research–A new researcher guide. *Shanlax International Journal of Education, 8*(4), 1–10. https://doi.org/10.34293/education.v8i4.3232

Huang, A. Y. Q., Lu, O. H. T., & Yang, S. J. H. (2023). Effects of artificial intelligence-enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education, 194*, Article 104684. https://doi.org/10.1016/j.compedu.2022.104684

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. https://doi.org/10.1016/j.caeai.2020.100001

Jacobsen, L. J., & Weber, K. E. (2023). *The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback*. OSF Preprints. https://doi.org/10.31219/osf.io/cr257

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, Article 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ..., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

Langley, P. (2019). An integrative framework for artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1). https://doi.org/10.1609/aaai.v33i01.33019670

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strwlan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024a). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence, 6*, Article 100221. https://doi.org/10.1016/j.caeai.2024.100221

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024b). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, Article 100213. https://doi.org/10.1016/j.caeai.2024.100213

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education, 31*, 186–214. https://doi.org/10.1007/s40593-020-00235-x

Lipnevich, A. A., Murano, D., Krannich, M., & Goetz, T. (2021). Should I grade or should I comment: Links among feedback, emotions, and performance. *Learning and Individual Differences, 89*, Article 102020. https://doi.org/10.1016/j.lindif.2021.102020

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*, Article 0028. https://doi.org/10.1038/s41562-016-0028

Madigan, D. J., & Kim, L. E. (2021). Does teacher burnout affect students? A systematic review of its association with academic achievement and student-reported outcomes. *International Journal of Educational Research, 105*, Article 101714. https://doi.org/10.1016/j.ijer.2020.101714

Magaldi, D., & Berler, M. (2020). Semi-structured interviews. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 4825–4830). Springer. https://doi.org/10.1007/978-3-319-24612-3_857

McCormick, K. (2015). *SPSS statistics for dummies*. John Wiley.

McGarrell, H., & Verbeem, J. (2007). Motivating revision of drafts through formative feedback. *ELT Journal, 61*(3), 228–236. https://doi.org/10.1093/elt/ccm030

Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology, 114*(8), 1743–1772. https://doi.org/10.1037/edu0000764

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Misiejuk, K., Kalissa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse. *Computers and Education: Artificial Intelligence, 6*, Article 100216. https://doi.org/10.1016/j.caeai.2024.100216

Pandero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*(5), Article 100416. https://doi.org/10.1016/j.edurev.2021.100416

Peters, K., & Halcomb, E. (2015). Interviews in qualitative research. *Nurse Researcher, 22*(4), 6–7. https://doi.org/10.7748/nr.22.4.6.s2

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Salcedo, J., & McCormick, K. (2020). *SPSS statistics* (4th ed.). John Wiley.

Schrader, C., & Kalyuga, S. (2020). Linking students' emotions to engagement and writing performance when learning Japanese letters with a pen-based tablet: An investigation based on individual pen pressure parameters. *International Journal of Human-Computer Studies, 135*, Article 102374. https://doi.org/10.1016/j.ijhcs.2019.102374

Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural. *Behaviour Research Methods, 38*(2), 262–279. https://doi.org/10.3758/BF03192778

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olsen, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., & Schuurmans, D. (2023). Foundation models for decision making: Problems, methods, and opportunities. *arXiv*. https://doi.org/10.48550/arXiv.2303.04129

Zawacki-Richter, O., Marin, 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*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zheng, Y., & Stewart, N. (2024). Improving EFL students' cultural awareness: Reframing moral dilemmatic stories with ChatGPT. *Computers and Education: Artificial Intelligence, 6*, Article 100223. https://doi.org/10.1016/j.caeai.2024.100223