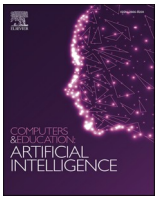
Computers and Education: Artificial Intelligence 6 (2024) 100199

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions

Jennifer Meyer [a,*], Thorben Jansen [a], Ronja Schiller [a], Lucas W. Liebenow [a], Marlene Steinbach [a], Andrea Horbach [b,c], Johanna Fleckenstein [b]

[a] Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118, Kiel, Germany
[b] University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany
[c] Fernuniversität Hagen, CATALPA, Universitätsstraße 47, 58097 Hagen, Germany

ARTICLE INFO

Keywords:
Secondary education
Improving classroom teaching
Applications in subject areas

ABSTRACT

Writing proficiency is an essential skill for upper secondary students that can be enhanced through effective feedback. Creating feedback on writing tasks, however, is time-intensive and presents a challenge for educators, often resulting in students receiving insufficient or no feedback. The advent of text-generating large language models (LLMs) offers a promising solution, namely, automated evidence-based feedback generation. Yet, empirical evidence from randomized controlled studies about the effectiveness of LLM-generated feedback is missing. To address this issue, the current study compared the effectiveness of LLM-generated feedback to no feedback. A sample of $N = 459$ upper secondary students of English as a foreign language wrote an argumentative essay. Students in the experimental group were asked to revise their text according to feedback that was generated using the LLM GPT-3.5-turbo. The control group revised their essays without receiving feedback. We assessed improvement in the revision using automated essay scoring. The results showed that LLM-generated feedback increased revision performance ($d = .19$) and task motivation ($d = 0.36$). Moreover, it increased positive emotions ($d = 0.34$) compared to revising without feedback. The findings highlight that using LLMs allows to create timely feedback that can positively relate to students' cognitive and affective-motivational outcomes. Future perspectives and the implications for research and practice of using LLM-generated feedback in intelligent tutoring systems are discussed.

1. Introduction

Writing is a complex and demanding cognitive process (Graham & Sandmel, 2011), including iterative cycles of planning, drafting, and revising (Flower & Hayes, 1981). Multiple cycles of high-quality revision are necessary for students to improve their writing. Staying motivated and focused throughout this effortful process can be challenging and creates the need for interventions that promote writing motivation (Bruning & Horn, 2000). Feedback from their teachers has been shown to be helpful (Graham et al., 2015) and motivating for students (Fong & Schallert, 2023; Graham, 2018). Similar effects were found for computer-generated feedback (Fleckenstein, Liebenow, & Meyer, 2023; Graham & Harris, 2017; Jansen et al., 2024; Lv et al., 2021; Ngo et al., 2022).

However, creating individual feedback for many students is difficult and time-consuming for teachers, especially on complex tasks such as essay writing (Hahn et al., 2021; Wambsganss et al., 2020; Zhu et al., 2020). Hence, teachers face the "bandwidth problem" (Wiley, 2006), which means that they can only support and guide a small number of students at a time; this is why it can be assumed that students do not receive feedback on a regular basis. To counter this problem, a large body of research has investigated the effectiveness of automated feedback that is based on automated writing evaluation (AWE; e.g., Bennett & Zhang, 2015; Crossley et al., 2022; Shermis, 2014). This body of research has found evidence that feedback supports students' writing even when the feedback is computer-based (for recent meta-analysis, see e.g., Ngo et al., 2022; Lv et al., 2021; Fleckenstein, Liebenow, & Meyer, 2023). Still, despite its effectiveness, AWE is

* Corresponding author.
E-mail address: jmeyer@leibniz-ipn.de (J. Meyer).

https://doi.org/10.1016/j.caeai.2023.100199
Received 18 August 2023; Received in revised form 20 December 2023; Accepted 24 December 2023
Available online 29 December 2023

limited in its applicability for independent use by classroom teachers in schools for several reasons: 1) AWE is task-specific, which means that the number of tasks available for automated scoring is limited, and the number of training opportunities when using AWE is insufficient (Nunes et al., 2022); 2) the generation of task-specific training data is cost- and time-intensive, as a large corpus of texts with human annotations is needed to train algorithms with sufficient reliability (Ramesh & Sanampudi, 2022; Rupp et al., 2019), which results in a small number of suitable corpora (Ramesh & Sanampudi, 2022); and 3) texts can only be scored automatically on the criteria that were rated in the training data sets, which limits flexibility in the evaluation of different text criteria and writing opportunities in the classroom. Accordingly, using automated feedback tailored individually to their students in their own classrooms for a large number of tasks was (until recently) out of reach for classroom teachers.

To improve this situation, recent developments in artificial intelligence (AI) that have increased the capabilities of large language models (LLMs) might be helpful, as they have the potential to automatically create individualized feedback on student performance at low cost for the educators (Kasneci et al., 2023) and without specific training data (Brown et al., 2020). However, there is a need for empirical evidence on the effectiveness of such automatically generated written feedback for student learning and affective-motivational learning outcomes. In this study, we assessed the effectiveness of LLM-generated feedback on an argumentative essay-writing task from the Test of English as a Foreign Language (TOEFL iBT®; ETS, 2023) in a sample of upper secondary students, focusing on revision performance, task motivation, students' positive emotional responses, and perceived feedback usefulness.

1.1. Feedback

Feedback is defined as information that is communicated to the learner to modify their thinking or behavior (e.g., Hattie & Timperley, 2007; Shute, 2008). Feedback provides support for learners on how to close the gap between their actual performance and a target performance (Biber et al., 2011; Hattie & Timperley, 2007), thereby aiming to improve learning (Shute, 2008), emotions, and motivation during a learning situation (Fong & Schallert, 2023). Feedback has been discussed from various theoretical perspectives that suggest that it is beneficial for learning and motivation (for an overview, see Thurlings et al., 2013; Panadero & Lipnevich, 2022). Based on these theories, feedback is considered effective when it is specific, timely (Shute, 2008), and when it answers the questions "Where am I going?", "How am I going?", and "Where to next?" (Hattie & Timperley, 2007).

There is also evidence showing that feedback is effective in the writing domain, with a body of empirical research showing that feedback is essential to support the revision process; this research is based on studies comparing the effects of feedback on writing to no feedback (Graham, 2018; Graham et al., 2023). In a systematic review of meta-analyses, Graham and Harris (2017) summarized (quasi-) experimental studies from meta-analyses on this topic. They found positive effect sizes of feedback from teachers or adults (7 primary studies, mean Cohen's $d = 0.87$), peer feedback (10 primary studies, mean $d = 0.77$), and computer feedback (5 primary studies, $d = 0.34$) on students' writing and revision performance.

1.1.1. How to provide feedback automatically: an overview of research on feedback based on AWE

The assessment provided by AWE systems is based on corpora of texts and analyses of their linguistic features (e.g., text length, mean word length, type-token ratio, syntactic complexity). Further, the texts need to be scored by human raters, that is, they need to be annotated regarding specific criteria related to writing performance, (e.g., scores of writing quality; Shermis, 2014, or argumentative elements; Crossley et al., 2022). Then, machine learning algorithms are used to predict the human scores recognizing patterns on the basis of the linguistic features in the training data. This allows evaluating new texts on the same task based on these patterns (see Ercikan & McCaffrey, 2022).

Prior research has shown that AWE can assess essays as accurately as trained humans can (Ramesh & Sanampudi, 2022) and, therefore, AWE systems are established in postsecondary admissions (e.g., TOEFL iBT®, Pearson Test of English) and learning systems (see Fleckenstein, Liebenow, & Meyer, 2023; Zhang, 2021). Studies have found that feedback that automatically adapts to students' strengths and weaknesses based on AWE scores can effectively foster drafting and revision performance (Fleckenstein, Liebenow, & Meyer, 2023; Horbach et al., 2022; Moore & MacArthur, 2016; Roscoe et al., 2014, Roscoe et al., 2019; Wade-Stein & Kintsch, 2004; Wilson, 2017; Wilson & Andrada, 2016) and productivity (Franzke et al., 2005; Huang & Wilson, 2021; Palermo & Thomson, 2018). However, it should be noted that AWE provides only the scoring of the text. The feedback message for students that goes beyond the mere scoring has to be created based on these scores and can be more or less effective (e.g., Mertens, Finn, & Lindner, 2022; Shute, 2008; Van der Kleij et al., 2015), and is not fully individualized to the specific texts.

So far, the effectiveness of feedback based on AWE has been evidenced by recent meta-analyses on the topic (Lv et al., 2021; Fleckenstein, Liebenow, & Meyer, 2023; Mohsen, 2022; Ngo et al., 2022; Zhai & Ma, 2023). This meta-analytic research shows that most studies on AWE feedback consider either the performance on a text revision or the performance on a new writing task as outcome measures of feedback effectiveness (Fleckenstein, Liebenow, & Meyer, 2023). These outcomes differ in their conceptualization: a successful revision describes a task improvement, and a successful application to a new task can be viewed as a learning outcome (Fleckenstein, Liebenow, & Meyer, 2023). Despite its effectiveness, as pointed out above, AWE has several shortcomings (e.g., they are task-specific [Nunes et al., 2022], cost and time-intensive to develop for a new task [Rupp et al., 2019; Ramesh & Sanampudi, 2022] and only applicable to criteria previously coded by human annotators [Ercikan & McCaffrey, 2022]) that limit its practical applicability in classrooms; these can be addressed by using LLMs to automatically generate feedback.

1.1.2. Using LLMs to generate feedback for students

LLMs, such as GPT-4 (OpenAI, 2023) or Claude (Bai et al., 2022), have emerged as a potential solution that widens the field of application for automated feedback to writing (UNESCO, 2023; Yang et al., 2023). LLMs are advanced AI systems that have been trained on vast amounts of textual data using the transformer architecture (Devlin, Chang, Lee, & Toutanova, 2018) and the attention mechanism (Vaswani et al., 2017) to analyze language patterns, which enables them to generate and manipulate human-like natural language. Compared to current AWE systems, one new feature of LLMs is that students and teachers can use them with natural language, allowing them to create automated feedback on their texts themselves with minimal further assistance (Kasneci et al., 2023). Natural language can be used to provide the LLM with concrete instructions (called "prompting") on how to create the feedback; using LLMs thus makes it possible to create feedback for each individual purpose, as prompts (i.e., "prompt engineering") can be refined until the feedback is appropriate for the respective feedback goal. For example, if the feedback created by LLMs is prompted to follow the feedback literature and provide timely answers to the students' questions "Where am I going?", "How am I going?", and "Where to next?" (Hattie & Timperley, 2007), it can be expected to foster students' revision performance. However, although this is great in theory, there is little evidence from randomized controlled studies of the effectiveness of such feedback that is generated using LLMs. Notably, LLM-generated feedback can have weaknesses that may influence feedback effectiveness. These have been discussed in the literature and can be due to the following aspects: a) the feedback can be false, b) the feedback can be biased, and c) the feedback is not transparent (for an overview of the strengths and weaknesses, see Chang et al., 2023). Still, using LLMs has great potential as it can enable teachers to provide their students with individual feedback (Kasneci et al., 2023), rewarding them

for their effort and thus increasing their motivation to stay engaged with revising their text.

Chang et al. (2023) reviewed the evidence currently available on the performance of LLMs, including GPT, across a variety of tasks. They reported that, for writing tasks, LLMs performed consistently across different genres, including argumentative and creative writing categories, as well as professional and informative texts, thereby showing that the writing capabilities of LLMs are general (Chia et al., 2023). Further, as reviewed by Chang et al. (2023), there is evidence that LLMs can successfully evaluate text quality without the use of reference texts and that they outperform most existing automated algorithms (Chen, Wang, Jiang, Shi, & Xu, 2023). In the educational context, a few studies have provided first empirical evidence that LLMs can be used to generate feedback for students: Dai et al. (2023) found that LLMs provided readable feedback on reports written by higher education students and that LLM- and instructor-generated feedback aligned regarding the polarity of feedback. Further, there is evidence showing that LLMs can augment learning by providing socioemotional support (Li & Xing, 2021). Further, research focusing on using LLM to provide feedback has focused mainly on perceptions of human evaluators outside of an authentic feedback context (e.g., Jacobsen & Weber, 2023; Steiss et al., 2023), providing preliminary evidence on the capabilities of LLMs as feedback providers. On the basis of these findings, we argue that feedback generated by LLMs can be a valuable aspect of modern writing instruction, even if it is not perfect and more research is needed on the determinants of its effectiveness.

1.1.3. The role of feedback in affective-motivational outcomes of the writing process

The writing process can be described as consisting of iterative cycles of planning, drafting, and revising (Flower & Hayes, 1981). Students in secondary school are often inexperienced with revision processes, due to a lack of feedback opportunities, which makes the task even more demanding. Moving through extensive cycles of drafting and revision creates unique motivational challenges (Bruning & Horn, 2000). Accordingly, feedback needs to be motivating and create positive emotional experiences in order for students to keep going and to apply continuous effort to the task at hand. Therefore, affective-motivational outcomes are important indicators of feedback effectiveness. For this reason, we considered them as relevant aspects of feedback effectiveness in this study.

An important aspect of students' task motivation (see Eccles & Wigfield, 2020; Troia et al., 2012) is students' intrinsic value beliefs towards tasks such as essay writing in English as a foreign language. Students who see value in engaging with a task because it is enjoyable and fun (i.e., intrinsic value) are more motivated to keep working on similar tasks. Feedback might help students stay motivated to actively engage with the task because it functions as a verbal reward. Rewards describe situations that have positive motivational properties from internal brain processes (Schultz, 2007). According to the cognitive evaluation theory (Deci & Ryan, 1985), feedback can function as a reward for students, raising feelings of competence by providing helpful information on how to succeed. Especially if the feedback is individualized according to the students' performance, it provides an external response and, thus, acknowledgment for the work that students put in when writing a draft. This assumption is backed by empirical evidence, showing that feedback from teachers can be a verbal reward (Hidi, Magnifico, & Renninger, 2023) and, for example, can enhance students' self-reported interest in a task (Deci et al., 1999).

In addition to motivation, students' emotions are central to the writing process (Lipnevich et al., 2021). Positive emotions (e.g., enjoyment, pride) can reinforce intrinsic motivation and engagement with a task, subsequently increasing learning success (e.g., Pekrun et al., 2023b; Eynde et al., 2007). Empirical evidence supports the function of feedback from teachers by showing positive effects of feedback on students' emotions (Lipnevich et al., 2021). Emotions can be influenced by task characteristics such as the presence of feedback, which can make students' experiences more personal and help them to persist with a task, even when the task (and the feedback) is computer-based (see Burleson & Picard, 2007).

Another aspect of feedback effectiveness is how the feedback is perceived by the students, in particular, whether feedback is perceived as useful to them. Highlighting the active role of the learner in feedback effectiveness (Winstone & Nash, 2023), prior research has shown that students need to perceive feedback as useful in order to elicit beneficial motivational, metacognitive, and cognitive responses (Harks et al., 2014; Mouratidis et al., 2010; Narciss, 2008; Panadero & Jonsson, 2013; Rakoczy et al., 2013, 2019; Shute, 2008).

In summary, although the feedback literature has been useful in describing and optimizing the creation of feedback from teachers, up until now, not much research has focused on the affective-motivational effects of computer-based and automated feedback (Camacho et al., 2021; Fleckenstein, Reble, et al., 2023). So far, empirical studies on feedback and affective-motivational outcomes have focused on corrective feedback on multiple-choice questions (Kuklick et al., 2023; Kuklick & Lindner, 2021, 2023), showing beneficial effects of affirmative feedback on students' affective-motivational outcomes but negative effects of corrective feedback. In the context of writing, only a small numbers of studies have investigated the motivational effects of AWE systems, showing beneficial effects for writing self-beliefs (e.g., Wilson & Roscoe, 2020). There is a need for further empirical research addressing the extent to which automated feedback, especially when generated by LLMs, can have motivational and emotional benefits.

1.1.3.1. The present study.

Prior research has shown that feedback can have cognitive and affective-motivational benefits for students, especially when it is individualized according to the students' prior performance (see Hattie & Timperley, 2007), and that it can also be effective when based on automated scoring algorithms e.g., Lv et al., 2021; Ngo et al., 2022; Fleckenstein, Liebenow, & Meyer, 2023). However, even with the availability of automated scoring techniques (e.g., AWE), creating automated feedback on complex tasks such as writing is costly, due to the need for large text corpora and high-quality annotations by extensively trained human raters, as well as the task-specificity of the currently available scoring algorithms (Rupp et al., 2019). LLMs enable researchers to tackle the issues that have hindered the widespread practical use of automated feedback on students' writing in the classroom (Kasneci et al., 2023; Yan et al., 2023). Thus, the focus of the current study is providing empirical evidence on the effectiveness of feedback created by LLMs, combining what is known from a large body of feedback literature with the new opportunities now available with LLMs.

In the present study, we used LLMs to create feedback based on principles from the feedback literature (i.e., prompting it to provide timely answers to the questions "Where am I going?", "How am I going?", and "Where to next?"; Hattie & Timperley, 2007). We hypothesize that such feedback can have cognitive benefits for students in their writing process, that is, that it can foster students' revision performance. Furthermore, we argue that due to the potential of feedback functioning as a verbal reward, feedback created by LLMs can support students' feelings of competence and acknowledgment of their efforts (Deci & Ryan, 1985; Lipnevich et al., 2021; Schultz, 2007), and thus foster students' motivation and elicit positive emotions. Investigating these assumptions, the present study addresses the following research questions:

RQ1 Can feedback generated by LLMs help students improve their revision performance?

RQ2 Can feedback generated by LLMs help students improve their performance in a new similar writing task?

RQ3 Can feedback generated by LLMs foster students' affective-motivational reactions after receiving feedback? More specifically, can LLM-generated feedback increase task motivation (RQ3a), elicit positive emotions (RQ3b) and do students report higher perceived usefulness of the LLM-generated feedback compared to a general instruction to revise (R3c)?

To investigate these questions, the current study compared the effectiveness of providing LLM-generated feedback to providing no feedback. We considered cognitive (i.e., revision performance in a writing task) as well as affective-motivational outcomes. We hypothesized that LLM-generated feedback would increase performance in text revision (RQ1) and performance in a new writing task (RQ2). We also hypothesized that the feedback would be beneficial for students' motivation (RQ3a) and would foster positive emotions (RQ3b). Highlighting the importance of the student perspective on feedback, we additionally assessed perceived feedback usefulness (RQ3c).

2. Materials and methods

2.1. Sample

We collected data from 552 upper secondary school students in Grade 10 between May 2023 and June 2023. We recruited the participants by sending letters to the principals of randomly drawn schools in the German federal state of Schleswig-Holstein whose schools had not participated in any studies within the past three months. The sample consisted of academic-track schools. Schools were told they would receive a report on the aggregated results of their school as compensation for participating. From the initial sample, we had to exclude 93 students due to technical difficulties that resulted in students not receiving any feedback to revise their first text. The final sample comprised $n = 459$ students (52% female; 40% male; 7.6% nonbinary or missing; mean age = 16.01 years; $n = 203$ in the feedback group, and $n = 256$ in the control group).

2.2. Study design

The data were collected in a 90-min classroom lesson. We brought tablets with keyboards to the classrooms. After a short introduction, students were asked to work on an argumentative writing task in English for 20 min. After working on the task, the students in the feedback group received LLM-generated feedback (see Section 2.3.3 for details); in the control group, the students were told that they would now have the opportunity to revise their text. Conditions were randomized within classrooms. After text revision, students rated their current emotions, their task motivation for future writing tasks, and the usefulness of the feedback. Students in both groups were then asked to work on the second task. We scored the texts from the first task, the revision, and the second task using an automated algorithm (see Section 2.3.2.2.2 Scoring Algorithm). A graphical illustration of the study design can be found in Fig. 1.

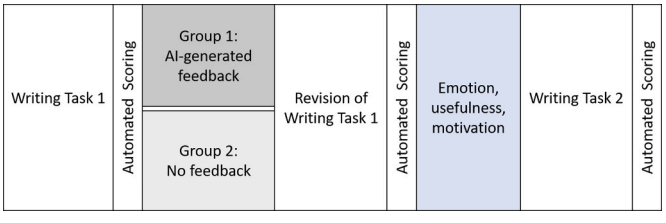


Fig. 1. Illustration of Study Procedure
Note. Students received two different writing tasks in randomized order.

2.3. Measurement

2.3.1. Instruments

2.3.1.1. Task motivation. We assessed students' task motivation in relation to the received feedback regarding future writing tasks in line with the intrinsic component of the situated expectancy-value theory (SEVT; Eccles & Wigfield, 2020), which captures students' anticipated enjoyment of a task in the future. We used a measure adapted from Busse et al. (2020) with four items (e.g., "Because of the feedback, I look forward to the next time I can work on similar writing tasks"). The internal consistency was acceptable, with $\alpha = 0.66$. Student responses were assessed on a Likert scale ranging from 1 (*does not apply at all*) to 6 (*fully applies*).

2.3.1.2. Emotions. We assessed affective reactions to feedback using epistemic emotions adapted from the German version (Vogl et al., 2018) of the Epistemically-Related Emotion Scales (Pekrun et al., 2017). We used eight items to capture the following emotions: surprise, curiosity, enjoyment, confusion, anxiety, frustration, boredom, and pride. Initially, we had planned to use both positive and negative emotions as outcome, but the reliability for negative emotions (boredom, frustration, and anxiety) was low with $\alpha = 0.35$. Accordingly, for the present analysis, we used only the emotions with a clearly positive valence (curiosity, enjoyment, and pride). The internal consistency was good with $\alpha = 0.62$. The items were assessed using the following instruction: We are interested in the feelings you experienced when reading the feedback you received for your first text. Please mark for each emotion, how intensively you experienced it by clicking the respective box. Student responses were assessed on a Likert scale ranging from 1 (*not at all*) to 5 (*very*).

2.3.1.3. Feedback usefulness. We assessed students' attitudes toward the feedback by asking students how useful they perceived the feedback to be. We used three items from Strijbos et al. (2021; e.g., "The feedback was helpful to me"). The internal consistency was good with $\alpha = 0.95$. Student responses were assessed on a Likert scale ranging from 1 (*does not apply at all*) to 7 (*fully applies*).

2.3.1.4. Background variables. We assessed several background variables in a questionnaire to compare the two groups regarding student gender, parental education, number of books at home, and last English grade.

2.3.2. Measures of writing performance

2.3.2.1. Writing task. Students were asked to write argumentative essays on two different tasks developed for the Test of English as a foreign language (TOEFL iBT®). Task A was: "Do you agree or disagree with the following statement? A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught. Use specific reasons and examples to support your answer." Task B was: "Do you agree or disagree with the following statement? Television advertising directed toward young children (aged two to five) should not be allowed. Use specific reasons and examples to support your answer." Each student was given both tasks, in a randomized order. Prior studies have shown that the tasks are comparable in difficulty (Keller et al., 2020; Rupp et al., 2019).

2.3.2.2. Automated evaluation

2.3.2.2.1. Text corpus and expert ratings. All texts were rated on rating scales developed for the Test of English as a foreign language (TOEFL iBT®). The text corpus is described in detail in Keller et al.

Table 1
Overview of mean and SD per group and significance test for differences between groups.

	M_{FG}	SD_{FG}	M_{CG}	SD_{CG}	Estimate	*SE*	*p* value
Demographics							
Gender [a]	0.58	0.49	0.56	0.50	0.02	0.04	.653
Education mother	1.48	0.74	1.59	1.00	−0.13	0.08	.088
Education father	1.52	0.89	1.69	1.13	−0.17	0.09	.062
Books at home	4.98	1.62	5.02	1.51	−0.03	0.09	.759
English grade [b]	2.64	1.13	2.64	1.21	−0.002	0.09	.985
First draft (Score 1)	2.47	0.97	2.45	0.95	0.02	0.10	.848
Outcomes							
Revision performance (Score 2)	2.87	0.92	2.70	0.88	0.19	0.09	.042
Motivation	3.11	1.02	3.51	1.14	0.36	0.11	.001
Positive emotions	1.95	0.75	1.68	0.79	0.34	0.12	.004
Feedback usefulness	3.72	1.56	1.64	1.27	1.19	0.08	<.001
New task (Score 3)	1.59	0.96	1.46	0.99	0.13	0.12	.259

Note. FG = feedback group, CG = control group, [a] 1 = female, 0 = male, [b] lower values = better achievement estimates report coefficients from regression analyses using the group (1 = feedback, 0 = control group) as predictor and the standardized outcome variable as criterion. Thus, effect sizes can be interpreted as Cohen's *d*. For the analysis that predicted Score 2, we conducted robustness checks using Score 1 as a covariate to control for prior performance ($estimate_{group} = 0.18$; $SE = 0.06$; $p = .013$). $^{*}p \leq .05$ $^{**}p \leq .01$ $^{***}p \leq .001$

(2020); the rater training and the rating process in Rupp et al. (2019). Our training data included texts from 2420 students (58.1% female, mean age = 17.7 years). Rigorously trained expert raters scored each text on a holistic scale ranging from zero (low quality) to five (high quality). The experts reached exact agreement on 62.5% of the texts and showed a quadratic weighted kappa of .67. When the experts did not agree on a text, a third expert with several years of experience as a master rater decided on the final score.

2.3.2.2.2. Scoring algorithm. On the basis of these annotated text data, we trained a machine learning algorithm to score students' writing (Text 1, Text 1 revision, Text 2) in a task-specific framework using a Java-based toolbox for free-text scoring based on natural language processing and machine learning (ESCRITO, Zesch & Horbach, 2018; see also Jansen et al., 2024). The scoring algorithm was built as follows. We used a support vector machine to train the scoring algorithm. In the broadest sense, a support vector machine is a linear regression in more than two dimensions. It is similar to linear regression in that the algorithm optimizes a prediction to have the minimum distance to the training data points. The main difference is that support vector machines use a higher dimensional surface, a hyperplane, as the decision boundary. To train the support vector machine, we initially calculated over 100 linguistic features to use them as predictors. The features included lexical information (word n-grams), structural information (part-of-speech n-grams), and features based on length and complexity. Second, we built the optimal hyperplane that separated the data points into six scores by minimizing the margin between the hyperplane and the closest data points. Third, we classified new texts using the trained algorithm. In a cross-validation setup of the training data, the algorithm reached an exact agreement with the human ratings on 43.9% of the texts and showed a quadratic weighted kappa of .76 with the human experts. Please note that we implemented the algorithm in the testing environment to score students' texts live, which allowed us to provide summative feedback immediately after the lesson. By doing so, we ensured that every student received feedback for their work, including those in the control group. This approach aligns with ethical practices by upholding the principle that all students deserve equal opportunities for learning and improvement.

2.3.3. LLM-generated feedback

The prompt to create feedback using LLMs was designed in line with feedback design principles from the feedback literature: it included the instruction to provide information for students on how to improve their text by providing hints and examples (Hattie & Timperley, 2007). Specifically, GPT was prompted that the feedback should include hints and examples for three aspects of text quality: structure, content, and language. The feedback should be given in tabular format, that is, in a structured way in order to decrease the cognitive load for students. It was also specified that the feedback should include short examples from the student text for individualization purposes. Further, the prompt included additional information on the student writers to which feedback should be given, noting that students would be foreign language learners in upper secondary school. The exact prompt given to GPT is shown in the Supplementary Material (Supplement 1). The following settings were used in GPT playground: model: GPT-3.5-turbo, temperature: 0, maximum length: 1800. Examples of feedback are shown in Fig. S2 in the supplements.

In the control group, students were given the opportunity to revise their text. They received the following message instead of individual feedback: "This is the feedback that you received for your first text. Please read your argumentative essay again and try to revise the text as best you can. Take sufficient time for your revisions." The ratings of feedback usefulness in the control group refer to this message.

2.4. Statistical analyses

We performed regression analyses using M*plus* (Version 8.2; Muthén & Muthén, 1998) to estimate group differences between the feedback and control groups. This means we used the group variable as the predictor. Additionally, to estimate differences in revision performance, we included the score on the first essay as a covariate to account for prior writing performance and we used the revision score as an outcome as a robustness check. We used the R package MplusAutomation to prepare the data and run the models (Hallquist & Wiley, 2018). The models were based on maximum likelihood estimation with robust standard errors (type = complex) to account for the dependency in the data due to the clustering in classrooms. To handle missing data within the questionnaire items, we used the full information maximum likelihood approach (FIML; see Enders, 2010). The amount of missing data ranged from 1% to 7%.

3. Results

3.1. Descriptives and bivariate correlations

Descriptive analyses showed that students did not differ in their initial writing performance (i.e., the score on their first draft) between the feedback and control groups (see Table 1). We also found no significant differences between the groups regarding demographics (i.e., number of books at home, parental education, gender, prior English grade; see Table 1). Generally, all students increased their performance

Table 2
Bivariate correlations.

	1	2	3	4	5	6	7	8	9	10
Demographics										
1 Gender [a]										
2 Education mother	−.01									
3 Education father	.03	.58**								
4 Books at home	.03	−.27***	−.26***							
5 English grade	−.08**	.14	.11	−.15*						
6 Score 1	.06*	−.16*	−.22**	.15**	−.23***					
Outcomes										
7 Positive emotions	−.01	.12	−.02	−.03	−.04	−.04				
8 Feedback usefulness	.02	.04	−.02	−.05	.08	−.05	.46***			
9 Motivation	.02	.05	−.05	−.01	.02	−.06	.40***	.46***		
10 Revision performance	.07*	−.19*	−.22**	.21***	−.26***	.65***	−.04	−.04	.03	
11 Score 3	.10***	−.11	−.11	.15*	−.34***	.25***	−.02	−.06	.01	.33***

Note. [a] 1 = female, 0 = male. *$p \leq .05$ **$p \leq .01$ ***$p \leq .001$

in the revision[1] (mean score Task 1 = 2.45, $SD = 0.96$; mean revision score = 2.77, SD = 0.90; $t = -8.502$; $df = 426$; $p < .001$; $d = 0.78$). Bivariate correlations between the variables can be found in Table 2.

3.2. Effectiveness of LLM-generated feedback

All results can be found in Table 1. For cognitive outcomes, our results show that students who received feedback generated by the LLM improved to a greater degree in the text revision compared to students in the control group (H1a). We found no differences between the groups for performance in a new subsequent writing task (H1b).

For the affective-motivational outcomes, we found that the feedback was beneficial for students' motivation to engage in subsequent similar tasks (H2), that is, students reported that they would enjoy working on similar tasks more after receiving the feedback. We found evidence that students reported more positive emotions (H3) after receiving LLM-generated feedback and working on the revision compared to the control group. Regarding perceived usefulness, students reported higher perceived usefulness of the feedback compared to that reported by students in the control group (H4). However, it should be noted that the mean reported usefulness in the feedback group was close to the mean of the scale (3.72 on a 1 to 7-point scale), indicating that the absolute usefulness of the feedback could be higher.

4. Discussion

The present study is to the best of our knowledge the first to provide empirical evidence for the effectiveness of LLM-generated feedback on student outcomes from a randomized controlled study in secondary education in the writing domain. Based on the prior literature showing the effectiveness of feedback in general (Hattie & Timperley, 2007; Graham & Sandmel, 2011), and for AWE writing outcomes (e.g., Lv et al., 2021; Fleckenstein, Liebenow, & Meyer, 2023; Mohsen, 2022; Ngo et al., 2022; Zhai & Ma, 2023) we hypothesized that feedback generated by LLMs would be effective in supporting students during text revision and would elicit beneficial affective-motivational processes in students that would help them stay motivated for the task during text revision (Cen & Zheng, 2024). Thus, the current study aimed to provide empirical evidence on the effectiveness of feedback created by LLMs, combining what is known from a large body of feedback literature with the new opportunities now available with LLMs. Specifically, our papers makes a contribution to the literature by providing empirical evidence on the potential that LLMs have for improving student writing with a tailored and efficient way to create personalized and immediate feedback for students.

For students' revision performance, we found, in line with our hypothesis, that students benefited from receiving LLM-generated feedback on an essay writing task compared to receiving no feedback. Thus, our evidence supports the much discussed idea (e.g., Cavalcanti et al., 2021; Kasneci et al., 2023; Kizilcec, 2023; Steiss et al., 2023; Yan et al., 2023) that LLMs can provide a way to implement findings from the feedback literature in the classroom, making it possible to provide effective feedback that can support students in revising their texts to a large number of students and thus increasing teachers' bandwidth regarding their individual feedback practice (Wiley, 2006). In other words, using LLMs to provide feedback allows teachers to provide individualized instruction to a large number of students at the same time. Notably, for revision performance, we found an effect size ($d = 0.19$) that was smaller compared to meta-analytic effect sizes previously found for feedback provided with AWE in secondary schools compared to a control group (Hedges' $g = 0.25$; Ngo et al., 2022; Hedges' $g = 0.27$ for revision tasks; see Fleckenstein, Liebenow, & Meyer, 2023). Still, this is a substantial effect considering the time and financial costs required to create feedback using AWE compared to LLMs, especially as LLMs can be adapted to new tasks, new evaluation criteria, or new insights from feedback research more easily than AWE-based feedback can.

However, meta-analytic research (e.g., Fleckenstein, Liebenow, & Meyer, 2023; Ngo et al., 2022) often includes both revision performance and performance on a new subsequent task as it integrates results from multiple study designs or populations (i.e. reporting mixed effect sizes for secondary and tertiary student populations). In the current study, we differentiated between the two types of tasks (i.e., revision and new task) and found significant differences only when considering the revision performance and not when considering the performance on a new task. Thus, we conclude that the feedback effects in this study were too small to elicit learning processes within one session; however, this might be expected as it seems unlikely that receiving one single feedback within just one learning session would result in sustainable learning gains. Instead, feedback might be most beneficial when provided more frequently. More research is needed that implements LLM-generated feedback over longer periods of time, as it can be assumed that the beneficial effects of feedback might unfold during subsequent learning opportunities (Fleckenstein, Liebenow, & Meyer, 2023). This calls for research that conducts longitudinal field experiments.

We found beneficial effects of the feedback that was generated using LLMs regarding motivation. The effects of LLM-generated feedback on motivation had larger effect sizes than the effects on performance. This result is particularly important for the feedback literature because the motivational effects of feedback have been assumed to be centrally linked to the social component of feedback from teachers and have not been discussed much for computer-based feedback so far (Fong &

[1] For this preliminary analysis, we used SPSS 28 and conducted a paired *t*-test; this means that, for this test, we did not account for hierarchies in the data and missing values.

Schallert, 2023). These positive effects align with the conceptualization of feedback as a reward (Deci & Ryan, 1985; Lipnevich et al., 2021; Schultz, 2007) and underscore the notion that receiving a timely reaction to a demanding performance is important to stay motivated, even when this reaction comes from a computer.

Regarding emotional reactions to the feedback, in line with the control-value theory (Pekrun, 2006; Pekrun et al., 2023a), our findings suggest that feedback generated using LLMs can foster positive control appraisals, which induce achievement emotions such as enjoyment and pride. Taken together, this supports previous findings that feedback can enhance students' emotional experiences and their perseverance with a task in a computer-based learning environment (Burleson & Picard, 2007). These results highlight that providing feedback might benefit students' learning, especially as it gives them a reason to stay motivated and to actively keep working on the task at hand (Wu & Schunn, 2023).

According to our results, the positive effect of feedback generated using LLMs can also be visible in students' perceptions of the feedback. We found that the LLM-generated feedback was more useful to students than receiving no feedback. The importance of this result is highlighted by the notion that feedback needs to be perceived as useful in order for it to elicit beneficial motivational, metacognitive, and cognitive responses (Harks et al., 2014; Mouratidis et al., 2010; Narciss, 2008; Panadero & Jonsson, 2013; Rakoczy et al., 2013, 2019; Shute, 2008). However, the mean score on the perceived usefulness scale was close to the mean of the scale, indicating that there might still be room for improvement regarding the perceived usefulness of LLM-generated feedback.

Taken together, based on our results, it seems plausible that making use of automatically generated feedback can allow for more practice opportunities for students to work on their writing skills, revise their texts multiple times, and receive feedback on each draft in less than a minute. Of course, to make the most of the potential of feedback that is generated using LLMs, teachers need to be familiar with the up- and downsides of such feedback (see Bogina, Hartman, Kuflik, & Shulner-Tal, 2022). Only then teachers can help students to actively work with the feedback but also inform them about potential biases and the probabilities of the feedback including incorrect information. Accordingly, the role of teachers needs to be considered in the design of educational technologies using LLMs (Kizilcec, 2023).

Thus, interventions that foster AI literacy in both teachers and students are required before such feedback is implemented on a larger scale. However, our study provides first evidence that feedback generated using LLMs promises new avenues for implementing feedback regularly during writing instruction, thereby releasing the pressure on teachers to provide complex feedback to all students and motivating students to keep revising their work.

4.1. Implications

Our results highlight the potential of using AI to create learning opportunities that are available with low cost for teachers and few practical constraints. As discussed in the literature, this can substantially relieve teachers' workload and provide students with more individualized learning opportunities, even in complex domains such as writing (e.g., Kasneci et al., 2023). We argue that even if the beneficial effect of one single LLM-generated feedback session is relatively small for revision performance, the effects of further feedback could cumulate for students because the automation allows for multiple feedback cycles and opportunities for practice. Accordingly, it can be hypothesized if LLM-generated feedback was implemented regularly in classroom practice, the overall effect could be much larger. This is where we see the potential and promise of LLM-generated feedback: No other type of teacher-based or even automated feedback allows for the implementation of feedback at scale for a variety of task types, competency levels, and topics.

Especially our findings regarding the affective-motivational benefits of the feedback can have important implications: Even though there might sometimes be errors in the AI-generated feedback, increases in motivation and emotion suggest that working with the feedback can enhance students' enjoyment when working on the task despite these limitations. Our study cannot rule out that these effects might be due to the novelty effects of working with such feedback for the first time (Clark, 1983). Accordingly, more research is needed to verify the possible implications of our findings. However, given the cost-effective implementation of LLM-based feedback, we believe that applying it in the classroom can have benefits, especially if the motivational effects would stay consistent after students have worked with LLM-generated feedback multiple times. Further research is needed in this area.

4.2. Limitations and future research

First, we would like to note that we provided feedback on only one task at one time in an experimental setting: further research is needed to investigate the potentially cumulative effects of the feedback in the long term and to consider follow-up measures that assess whether the feedback only improves performance or whether it triggers learning processes that students can benefit from later on. Furthermore, future studies need to investigate whether the regular use of AI-based feedback systems can also have implications for real-life outcomes such as grades or GPA. Regarding the finding that receiving feedback from LLMs once during the revision was not enough to benefit writing performance in a subsequent writing task, we argue that the cumulative effects need to be investigated in order to understand whether receiving such feedback can increase learning to a substantial degree.

Second, although we tried to align the LLM-generated feedback with the empirical feedback literature, it is still possible that the quality of the feedback varied across students. We did not consider the quality of the feedback in our investigation as we were interested in the mean effects; at this point, we cannot rule out the possibility that some of the feedback might have been of insufficient quality. Due to the nature of LLMs, we could not completely control the quality; feedback generated by LLMs will always vary more or less. Students need to be taught to understand the circumstances under which the feedback is created, about its challenges and potential, and strategies how to use LLM-based tools (Tseng & Warschauer, 2023; Warschauer et al., 2023). If students have basic AI literacy regarding LLMs, they will be able to handle feedback situations that might be less than optimal. Similarly, students can react differently to different feedback depending on their learning prerequisites, personalities, and learning goals (Panadero & Lipnevich, 2022). Research is needed that takes into account how different students might benefit from different types of LLM-generated feedback.

Third, our study cannot provide insights into the causality of why we found effects for text revision and affective-motivational outcomes. We can only show that students appreciated the feedback, as evidenced by motivational factors in response to the feedback, and that there were cognitive benefits, as evidenced by the increased revision performance compared to that of the control group.

Fourth, we did not explicitly inform the participating students beforehand that the feedback would be provided by LLMs. We do not know whether the students had implicit hypotheses about the feedback source, which could potentially have triggered negative or positive attitudes towards AI-based feedback and could have influenced our results. Research is needed to understand how students perceive AI-based feedback when the source is known or unknown and how this might affect their active engagement with the feedback. Additionally, in the present study we cannot compare the effectiveness of LLM-generated feedback with teacher-based feedback. Studies are needed that compare students receiving LLM-generated feedback with students receiving teacher feedback to fully assess the effectiveness of LLM-generated feedback. It is an important limitation of our study that we cannot attribute the results to the quality of the LLM-generated feedback but instead it is possible that our results are due to the presence of any feedback regardless of its quality.

Fifth, the feedback prompt we used in this study was based on the feedback literature (Hattie & Timperley, 2007). However, it is still not fully clear what design features determine the effectiveness of feedback (especially considering different students and their preferences; Panadero & Lipnevich, 2022; Winstone & Nash, 2023). Accordingly, future research should closely investigate the role of feedback characteristics within AI-generated feedback. For example, LLMs can be prompted to provide feedback in tabular format, as we did in this study, which might reduce the cognitive load for the students working with the feedback compared to the cognitive load of feedback presented as a continuous text, as such elements of feedback or task design might impact extraneous cognitive load (e.g., DeLeeuw & Mayer, 2008; Redifer et al., 2021). Increased extraneous cognitive load is known to be detrimental to complex task performance (Sweller, 2011). Similarly, feedback effects can vary depending on other feedback characteristics (see Mertens et al., 2022), for example evaluative or elaborate feedback components. Such questions of effective feedback design could potentially be addressed more economically with studies using AI-generated feedback than with traditional feedback studies. Similarly, more work needs to be done on the successful prompting of different LLMs to generate the most beneficial feedback for students, also taking account of their individual learning prerequisites.

Sixth, the choice of scoring algorithm could have influenced the results of our study. There is research showing that transformer-based neural networks can provide superior accuracy (see Ormerod, Malhotra, & Jafari, 2021; Ludwig et al., 2021). However, different evaluation metrics are known to produce different results for different application contexts (Doewes et al., 2023) and thus, accuracy might not be the only relevant performance indicator in a specific use-case. More research is needed to understand the implications of using different algorithms for specific applied use cases in educational research. For example, the distribution of scores might be an important aspect that needs to be assessed in balance with scoring accuracy to determine the most appropriate scoring algorithm, especially when students' individual development is to be assessed.

Finally, in the current study, we focused only on mean differences between our experimental groups. However, considering questions of fairness and equity in education, it is important to investigate whether AI-generated feedback benefits some students more than others. Accordingly, future studies are needed to address whether AI-generated feedback is equally effective for different student groups (e.g., students at the early or later stages of their learning process, students of different genders or with different backgrounds), also considering potential algorithmic biases of AI-generated feedback towards certain groups of students (e.g., Dieterle, Dede, & Walker, 2022; Schramowski et al., 2022; Li et al., 2023).

4.3. Conclusion

Our study contributes to the literature by providing empirical evidence on the potential of LLMs in education (Kasneci et al., 2023), specifically as a feedback giver. The present study is the first to provide experimental evidence of the effectiveness of LLM-generated feedback for student outcomes in the writing domain. Even though more research is needed in this field, especially regarding the lasting benefits of AI-based feedback beyond the novelty phase, our results highlight the potential that LLMs hold for providing individualized instruction to all students, thereby yielding cognitive and affective-motivational benefits, while, at the same time, allowing low-cost and timely implementation in the classroom. LLM-generated feedback, even considering its weaknesses (e.g., Zhuo, Huang, Chen, & Xing, 2023), could together with other use-cases, transform the field of writing instruction, allowing for more practice opportunities that can lead to growth in student competencies. Our study provides a starting point by highlighting the potential of LLMs to provide feedback. On the basis of these results, there is a need to prepare teachers and students to make use of the potential of these cutting-edge technologies.

5. Statements on open data and transparency

This study was not preregistered. All data and syntax that can be used to reproduce the results of this study can be found at https://osf.io/w7vh9/. This study was reviewed by the Ministry of General Education and Vocational Training, Science, Research and Culture in Schleswig-Holstein and was approved by the ethics committee at the Leibniz-Institute for Science and Mathematics Education.

CRediT authorship contribution statement

Jennifer Meyer: Conceptualization, Formal analysis, Writing – original draft. **Thorben Jansen:** Conceptualization, Writing – review & editing. **Ronja Schiller:** Investigation, Validation, Writing – review & editing. **Lucas W. Liebenow:** Investigation, Validation, Writing – review & editing. **Marlene Steinbach:** Investigation, Validation, Writing – review & editing. **Andrea Horbach:** Formal analysis, Methodology, Software, Writing – review & editing. **Johanna Fleckenstein:** Conceptualization, Funding acquisition, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-4 from OpenAI in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Gráinne Newcombe for language editing. This work was funded by the German Federal Ministry of Education and Research grant number 01JG2104.

List of Acronyms

LLMs	Large language models
TOEFL iBT®	Test of English as a foreign language
ETS	Educational Testing Service
AI	Artificial Intelligence
GPT	Generative pre-trained transformer
AWE	Automated Writing Evaluation

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.caeai.2023.100199.

References

Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing* (pp. 142–173). Routledge. https://doi.org/10.4324/9781315871493.

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series, 2011*(1), i–99. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Bogina, V., Hartman, A, Kuflik, T., & Shulner-Tal, A. (2022). Educating Software and AI Stakeholders About Algorithmic Fairness. *Accountability, Transparency and Ethics. Int J Artif Intell Educ, 32*(3), 808–833. https://doi.org/10.1007/s40593-021-00248-0

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Bruning, R., & Horn, C. (2000). Developing motivation to write. *Educational Psychologist, 35*(1), 25–37. https://doi.org/10.1207/S15326985EP3501_4

Burleson, W., & Picard, R. W. (2007). Gender-specific approaches to developing emotionally intelligent learning companions. *IEEE Intelligent Systems, 22*(4), 62–69. https://doi.org/10.1109/MIS.2007.69

Busse, V., Krause, U. M., Parr, J., & Schubert, P. (2020). Developing secondary students' writing skills: Affective and motivational effects of a feedback intervention with learners of English as a foreign language. *Classroom Observation: Researching Interaction in English Language Teaching*, 245–265.

Camacho, A., Alves, R. A., & Boscolo, P. (2021). Writing motivation in school: A systematic review of empirical research in the early twenty-first century. *Educational Psychology Review, 33*(1), 213–247. https://doi.org/10.1007/s10648-020-09530-4

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

Cen, Y., & Zheng, Y. (2024). The motivational aspect of feedback: A meta-analysis on the effect of different feedback practices on L2 learners' writing motivation. Assessing Writing, 59, 100802. https://doi.org/10.1016/j.asw.2023.100802.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., … Xie, X. (2023). *A survey on evaluation of large language models*. arXiv preprint arXiv:2307.03109v5.

Chen, Y., Wang, R., Jiang, H., Shi, S., & Xu, R. (2023). *Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study*. arXiv preprint arXiv:2304.00723.

Chia, Y. K., Hong, P., Bing, L., & Poria, S. (2023). *Instructeval: Towards holistic evaluation of instruction-tuned large language models*. arXiv preprint arXiv:2306.04757.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing, 54*. https://doi.org/10.1016/j.asw.2022.100667. Article 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. *Preprint*. https://doi.org/10.35542/osf.io/hcgz

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. https://doi.org/10.1037/0033-2909.125.6.627

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer US. https://doi.org/10.1007/978-1-4899-2271-7

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. 1810.04805v2.

Dieterle, E., Dede, C., & Walker, M. (2022). *The cyclical ethical effects of using artificial intelligence in education* (pp. 1–11). AI & society.

Doewes, A., Kurdhi, N., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th international Conference on educational data mining, EDM 2023* (pp. 103–113). International Educational Data Mining Society (IEDMS).

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://psycnet.apa.org/doi/10.1016/j.cedpsych.2020.101859.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement, 59*(3), 272–287. https://doi.org/10.1111/jedm.12332

Eynde, P. O. 't, Corte, E. de, & Verschaffel, L. (2007). Students' emotions. In P. A. Schultz, & R. Pekrun (Eds.), *Emotion in education* (pp. 185–204). Elsevier. https://doi.org/10.1016/B978-012372545-5/50012-5.

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023a). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, Article 1162454. https://doi.org/10.3389/frai.2023.1162454

Fleckenstein, J., Reble, R., Meyer, J., Jansen, T., Liebenow, L. W., Möller, J., & Köller, O. (2023b). Digitale Schreibförderung im Bildungskontext: Ein systematisches Review. [Digital Writing Instruction in the educational context: A systematic review. In K. Scheiter, & I. Gogolin (Eds.), *Bildung für eine digitale Zukunft [Education for a digital future* (pp. 3–25). Wiesbaden: Springer Fachmedien Wiesbaden.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition & Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Fong, C. J., & Schallert, D. L. (2023). "Feedback to the future": Advancing motivational and emotional perspectives in feedback research. *Educational Psychologist*, 1–16. https://doi.org/10.1080/00461520.2022.2134135

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33*(1), 53–80. https://doi.org/10.2190/DH8F-QJWM-J457-FQVB

Graham, S. (2018). A revised writer (s)-within-community model of writing. *Educational Psychologist, 53*(4), 258–279. https://doi.org/10.1080/00461520.2018.1481406

Graham, S., & Harris, K. R. (2017). Evidence-based writing practices: A meta-analysis of existing meta-analyses. In *Design principles for teaching effective writing* (pp. 13–37). Brill. https://doi.org/10.1163/9789004270480_003.

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

Graham, S., Kim, Y.-S., Cao, Y., Lee, J.w., Tate, T., Collins, P., … Olson, C. B. (2023). A meta-analysis of writing treatments for students in grades 6–12. *Journal of Educational Psychology, 115*(7), 1004–1027. https://doi.org/10.1037/edu0000819.

Graham, S., & Sandmel, K. (2011). The process writing approach: A meta-analysis. *The Journal of Educational Research, 104*(6), 396–407. https://doi.org/10.1080/00220671.2010.488703

Hahn, M. G., Navarro, S. M. B., La Fuente Valentin, L. de, & 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

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology, 34*(3), 269–290. https://doi.org/10.1080/01443410.2013.785384

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hidi, S. E., Magnifico, A., & Renninger, K. A. (2023). Students developing as writers: How and why interest makes a difference. In R. Horowitz (Ed.), *The Routledge Handbook of International Research on Writing* (2nd Edition, pp. 477–492). New York: Routledge.

Horbach, A., Laarmann-Quante, R., Liebenow, L., Jansen, T., Keller, S., Meyer, J., … Fleckenstein, J. (2022). Bringing automatic scoring into the classroom–measuring the impact of automated analytic feedback on student writing performance. In *Swedish language technology conference and NLP4CALL* (pp. 72–83).

Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition, 62*, Article 102675. https://doi.org/10.1016/j.compcom.2021.102675

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://osf.io/cr257.

Jansen, T., Meyer, J., Fleckenstein, J., Horbach, A., Keller, S., & Möller, J. (2024). Individualizing goal-setting interventions using automated writing evaluation to support secondary school students' text revisions. *Learning and Instruction, 89*, 101847. https://doi.org/10.1016/j.learninstruc.2023.101847.

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

Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., & Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing, 48*, Article 100700.

Kizilcec, R. F. (2023). To advance AI use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education*, 1–8. https://doi.org/10.1007/s40593-023-00351-4

Kuklick, L., Greiff, S., & Lindner, M. A. (2023). Computer-based performance feedback: Effects of error message complexity on cognitive, metacognitive, and motivational outcomes. *Computers & Education, 200*(1), Article 104785. https://doi.org/10.1016/j.compedu.2023.104785

Kuklick, L., & Lindner, M. A. (2021). Computer-based knowledge of results feedback in different delivery modes: Effects on performance, motivation, and achievement emotions. *Contemporary Educational Psychology, 67*, Article 102001. https://doi.org/10.1016/j.cedpsych.2021.102001

Kuklick, L., & Lindner, M. A. (2023). Affective-motivational effects of performance feedback in computer-based assessment: Does error message complexity matter? *Contemporary Educational Psychology, 73*, Article 102146. https://doi.org/10.1016/j.cedpsych.2022.102146

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

Li, T., Reigh, E., He, P., & Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching, 60*(6), 1385–1389. https://doi.org/10.1002/tea.21867

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

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych, 3*(4), 897–915.

Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *The Asia-Pacific Education Researcher, 30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6

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.

Mohsen, M. A. (2022). Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis. *Journal of Educational Computing Research, 60*(5), 1253–1276. https://doi.org/10.1177/07356331211064066

Moore, N., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research, 8*(1), 149–175. https://doi.org/10.17239/jowr-2016.08.01.05

Mouratidis, A., Lens, W., & Vansteenkiste, M. (2010). How you provide corrective feedback makes a difference: The motivating role of communicating in an autonomy-supporting way. *Journal of Sport & Exercise Psychology, 32*(5), 619–637. https://doi.org/10.1123/jsep.32.5.619

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 125–143). Routledge.

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2022). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2022.2096642

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning, 38*(2), 599–620.

Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). *Automated essay scoring using efficient transformer-based language models*. arXiv preprint arXiv:2102.13136.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of Self-Regulated Strategy Development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255–270. https://doi.org/10.1016/j.cedpsych.2018.07.002

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, 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

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., … Vispoel, W. P. (2023a). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology, 124*(1), 145. https://doi.org/10.1037/a0013383

Pekrun, R., Marsh, H. W., Suessenbach, F., Frenzel, A. C., & Goetz, T. (2023b). School grades and students' emotions: Longitudinal models of within-person reciprocal effects. *Learning and Instruction, 83*, 101626.

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra. (2017). Measuring emotions during epistemic activities: The epistemically-related emotion scales (EES). *Cognition & Emotion, 31*, 1268–1276. https://doi.org/10.1080/02699931.2016.1204989

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63–73. https://doi.org/10.1016/j.learninstruc.2013.03.002

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction, 60*, 154–165. https://doi.org/10.1016/j.learninstruc.2018.01.004

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

Redifer, J. L., Bae, C. L., & Zhao, Q. (2021). Self-efficacy and performance feedback: Impacts on cognitive load during creative thinking. *Learning and Instruction, 71*, Article 101395.

Roscoe, R. D., Allen, L. K., & McNamara, D. S. (2019). Contrasting writing practice formats in a writing strategy tutoring system. *Journal of Educational Computing Research, 57*(3), 723–754. https://doi.org/10.1177/0735633118763429

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Research Report Series, 2019*(1), 1–23.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258–268. https://doi.org/10.1038/s42256-022-00458-8

Schultz, W. (2007). Reward. *Scholarpedia, 2*(3), 1652.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Steiss, J., Powell Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., & Warschauer, M. (2023). Comparing the quality of human and ChatGPT feedback on students' writing. *OSF Preprints*. https://doi.org/10.35542/osf.io/ty3em

Strijbos, J.-W., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the feedback perceptions questionnaire. *Studies In Educational Evaluation, 68*, Article 100980. https://doi:10.1016/j.stueduc.2021.100.

Sweller, J. (2011). Cognitive load theory. In , *(Vol. 55,. Psychology of learning and motivation* (pp. 37–76). Academic Press.

Thurlings, M., Vermeulen, M., Bastiaens, T., & Stijnen, S. (2013). Understanding feedback: A learning theory perspective. *Educational Research Review, 9*, 1–15. https://doi.org/10.1016/j.edurev.2012.11.004

Troia, G. A., Shankland, R. K., & Wolbers, K. A. (2012). Motivation research in writing: Theoretical and empirical considerations. *Reading & Writing Quarterly, 28*(1), 5–28. https://doi.org/10.1080/10573569.2012.632729

Tseng, W., & Warschauer, M. (2023). AI-Writing tools in education: If you can't beat them, join them. In *Journal of China computer-assisted language learning* (Vol. 0). Online First.

UNESCO. (2023). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Vogl, E., Pekrun, R., & Muis, K. R. (2018). Validierung einer deutschsprachigen Skala zur Messung epistemischer Emotionen. In G. Hagenauer, & T. Hascher (Eds.), *Emotionen und Emotionsregulation in der Schule und Hochschule* (pp. 259–272). Waxmann.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333–362. https://doi.org/10.1207/s1532690xci2203_3

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14).

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). *The affordances and contradictions of AI-generated text for second language writers*. Available at SSRN.

Wiley, D. A. (2006). Learning objects in public and higher education. In *Innovations in instructional technology* (pp. 1–9). Routledge.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing, 30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality. In J. Keengwe, Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Advances in Higher Education and Professional Development. Handbook of Research on Technology Tools for Real-World Skill Development (S. 679–704). IGI Global*.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87–125.

Winstone, N. E., & Nash, R. A. (2023). Toward a cohesive psychological science of effective feedback. *Educational Psychologist*, 1–19. https://doi.org/10.1080/00461520.2023.2224444

Wu, Y., & Schunn, C. D. (2023). Passive, active, and constructive engagement with peer feedback: A revised model of learning from peer feedback. *Contemporary Educational Psychology, 73*, Article 102160. https://doi.org/10.1016/j.cedpsych.2023.102160

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

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 literature review. *arXiv preprint arXiv:2303*, Article 13379.

Zesch, T., & Horbach, A. (2018, May). Escrito-an nlp-enhanced educational scoring toolkit. In *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research, 61*(4), 875–900. https://doi.org/10.1177/07356331221127300

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning, 1*(1), 170–176. https://doi.org/10.1515/jccall-2021-2007

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, Article 103668. https://doi.org/10.1016/j.compedu.2019.103668

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness. *Reliability and Toxicity*. arXiv preprint arXiv:2301.12867.