Computers & Education 237 (2025) 105385

Contents lists available at ScienceDirect

Computers & Education

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

Self-assessment accuracy in the age of artificial Intelligence: Differential effects of LLM-generated feedback

Lucas W. Liebenow [a,*], Fabian T.C. Schmidt [b], Jennifer Meyer [a,c], Johanna Fleckenstein [b]

[a] Educational Research and Educational Psychology, Leibniz Institute for Science and Mathematics Education, Germany
[b] Applied Educational Science, University of Hildesheim, Germany
[c] Department for Teacher Education, University of Vienna, Austria

ARTICLE INFO

Keywords:
Self-assessment accuracy
Feedback
Large language models
Artificial intelligence
Monitoring accuracy

ABSTRACT

Feedback is a promising intervention to foster students' self-assessment accuracy (SAA), but the effect can vary depending on students' initial skill levels or prior performance. In particular, lower-performing students who are less accurate might benefit more from feedback in terms of SAA. To deepen our understanding, the present study investigated the mechanism and dependencies of feedback effects on SAA in the realm of large language models (LLMs). Within a randomized control experiment, we examined the effect of LLM-generated feedback on SAA by considering students' initial performance and initial SAA as potential moderators. A sample of $N = 459$ upper secondary students wrote an argumentative essay in English as a foreign language and revised their text. After finishing their first draft (pretest) and revision (posttest) of the draft, students self-assessed their writing performance. Students in the experimental group received GPT-3.5-turbo-generated feedback on their first draft during their revision. In the control group, students could revise their text without feedback. Our results indicated no significant main effect of LLM-generated feedback on students' SAA. Furthermore, we found a significant interaction effect between feedback and students' pretest SAA on SAA changes, indicating that lower-calibrated students improved their SAA with feedback more than students with similar pretest SAA and without feedback. Exploratory analyses revealed that students with higher pretest SAA did not improve their SAA with feedback and decreased their SAA. We discuss this nuanced evidence and draw implications for research and practice using LLM-generated feedback in education.

1. Introduction

With recent developments in artificial intelligence (AI), technology-based learning has become increasingly important (Fleckenstein et al., 2024; Kasneci et al., 2023; Yan et al., 2024). To succeed in these educational settings, students have to self-regulate their learning processes: They need to effectively manage their time, set personal learning goals, and monitor their progress (Broadbent & Poon, 2015; Schwam et al., 2021). To ensure effective self-regulation, students need to self-assess their abilities and performance

* Corresponding author. Educational Research and Educational Psychology, Leibniz Institute for Science and Mathematics Education, Olshausenstrasse 62, Kiel, Schleswig-Holstein, D-24118, Germany.
E-mail address: liebenow@leibniz-ipn.de (L.W. Liebenow).

https://doi.org/10.1016/j.compedu.2025.105385
Received 10 January 2025; Received in revised form 8 May 2025; Accepted 18 June 2025
Available online 19 June 2025

accurately, as accurate self-assessment enables students to set attainable goals and to adequately adjust their learning strategies (Baas et al., 2015; De Bruin & van Gog, 2012; Koriat, 2012; Panadero, 2017). Self-assessment accuracy (SAA) can be described as the correspondence between self-assessed performance and actual performance outcomes (Panadero et al., 2016; Schraw, 2009). Research has repeatedly found that students tend to be inaccurate in their self-assessment and usually overestimate their performance (Falchikov & Boud, 1989; León et al., 2023), which has prompted calls for effective interventions to support students' SAA (Brown & Harris, 2014; León et al., 2023).

Providing feedback to students is a promising intervention, as it provides information that students can use to guide their self-assessment (Butler & Winne, 1995; Panadero & Alqassab, 2019). Research over the last three decades has indicated that feedback has positive effects on students' SAA (e.g., Braumann et al., 2024; Lichtenstein & Fischhoff, 1980; Nietfeld et al., 2006). In particular, feedback seems to be beneficial for lower-performing students, who tend to be less accurate in their self-assessment than higher-performing students and, therefore, need more support (Liebenow et al., 2024; Nederhand et al., 2019). This implies that students with lower SAA — who, in turn, typically also perform worse (Kruger & Dunning, 1999) — may also benefit more from feedback. This is not only because they have more room for improvement, but also because, similar to lower-performing students, they are more likely to benefit from external, valid information such as feedback (Thiede et al., 2010). This emphasizes that the effect of feedback can vary depending on individual differences, that is, on students' levels of SAA and performance.

Providing feedback can be time-consuming for teachers, particularly when it comes to complex writing tasks (Fleckenstein et al., 2023). This challenge can be addressed by using AI, in particular, large language models (LLMs), which can provide real-time, individualized feedback to a large number of students simultaneously (Yan et al., 2024). Initial studies on this topic have presented promising evidence. For instance, it has been demonstrated that GPT-generated feedback can have a positive effect on writing performance, emotions, and motivation (e.g., Meyer et al., 2024). However, research is needed to determine whether the potential benefits of LLM-generated feedback also apply to SAA. Differences between LLM-generated and human feedback—such as lower feedback accuracy (Steiss et al., 2024) or less valid performance scores (Nilsson & Tuvstedt, 2023)—may influence its effect on SAA. For instance, evaluative feedback with performance scores is particularly beneficial for SAA (Liebenow et al., 2024). Therefore, the present study investigated whether previous findings concerning feedback effects on SAA can be extended to LLM-generated feedback. To do so, we examined the effect of LLM-generated feedback on SAA in the context of a writing task in a randomized control experiment. Moreover, to provide a more nuanced understanding of how feedback affects SAA, we examined whether its effectiveness depends on students' initial performance and SAA, thereby following the idea that feedback serves a compensatory function for students who struggle to monitor and evaluate their own performance (Panadero et al., 2016).

1.1. Self-assessment accuracy

Self-assessment encompasses different methods and techniques that allow students to monitor and self-assess their learning progress and performance (Panadero et al., 2016). SAA—also known as calibration accuracy (e.g., Bol et al., 2005; Hacker & Bol, 2019) or metacognitive monitoring accuracy (e.g., De Bruin et al., 2017; Rawson and Dunlosky, 2007)—describes the degree to which students' self-assessed performance corresponds to their actual performance outcome (Panadero et al., 2016; Schraw, 2009). Research on self-assessment is predominantly informed by self-regulated learning theories, which conceptualize self-assessment as a key metacognitive process within the learning process (Butler & Winne, 1995; Nelson & Narens, 1990; Panadero et al., 2016; Zimmerman, 2000). An accurate self-assessment allows students to form a realistic understanding of their own learning processes and outcomes (Nelson & Narens, 1990). This, in turn, enhances self-regulated learning by helping students establish achievable goals, monitor their progress more precisely, and adapt their strategies appropriately (Panadero, 2017; Zimmerman, 1989). In fact, engaging in self-assessment has been shown to promote students' self-regulatory skills by encouraging reflection and accurate self-monitoring (Panadero & Alonso-Tapia, 2013; Panadero et al., 2017).

Following the crucial role of SAA within self-regulated learning processes, the link between SAA and academic performance becomes evident (Nelson & Narens, 1990). Students who self-assess their performance more accurately are more likely to make adequate adaptions to their current learning behavior and, therefore, perform better than less accurate students (Dunlosky & Rawson, 2012; Thiede et al., 2010). In line with this, Thiede et al. (2003) found that students with higher SAA regarding their text comprehension were able to more adequately identify the texts they needed to restudy and subsequently achieved higher performance than students with lower SAA. Taken together, this underscores and emphasizes the importance of SAA within educational research and practices, which has been described as "a sine qua non for effective learning" (Black & Wiliam, 1998, p. 26).

1.2. Self-assessment accuracy and feedback

It has frequently been observed that students are prone to inaccuracy in their self-assessment (Falchikov & Boud, 1989; León et al., 2023), which underscores the need for effective interventions to support students' SAA. This phenomenon can be elucidated through the lens of the cue-utilization framework (Koriat, 1997; Koriat et al., 2008). According to this framework, students use cues (i.e., information about their learning and performance) as an orientation for their self-assessment. These cues can be of varying quality, and Koriat (1997) distinguished between diagnostic cues (i.e., information that predicts actual performance) and nondiagnostic cues (information that is nonpredictive for actual performance). For example, students may falsely assess their text comprehension based on their reading speed, which is usually a nondiagnostic cue, and, therefore, might assess their text comprehension inaccurately. In contrast, when students receive, for instance, teachers' feedback for revising their text that includes diagnostic cues, they can use this high-quality feedback as an orientation for their self-assessment during the revision process, which can, in turn, enhance the accuracy

of their self-assessment. Scholars argued that students tend to be inaccurate in their self-assessment, as students frequently use non-diagnostic cues or due to a lack of diagnostic cues (De Bruin & Van Merriënboer, 2017; Händel et al., 2020; Koriat et al., 2008). In other words, students need to be provided with valid diagnostic cues that support their SAA.

This facilitates feedback that can communicate diagnostic cues during students' self-assessment (Butler & Winne, 1995; Panadero et al., 2016; Thiede et al., 2010). Feedback has been described as an inherent catalyst for self-assessment processes (Butler & Winne, 1995), and research over the last three decades has indicated that feedback has positive effects on students' SAA (e.g., Arkes et al., 1987; Braumann et al., 2024; Labuhn et al., 2010; Nederhand et al., 2019). Recently, Liebenow et al. (2024) conducted a meta-analysis on the effect of feedback on students' SAA and found a significant effect ($g = 0.34$), indicating the promising potential of feedback to foster SAA. Possible pathways that explain the effect of feedback on SAA are that students correct their self-assessment using diagnostic cues from the feedback and thereby become more accurate. This, in turn, can facilitate students' lifelong learning and academic achievement, which underscores the importance of interventions to support students' self-assessment processes (Panadero et al., 2016; Brown & Harris, 2014).

1.3. Differential effects of feedback

In general, lower-performing students tend to be less accurate in their self-assessment than higher-performing students (Ehrlinger et al., 2008; Kruger & Dunning, 1999), which in turn might both influence the effect of feedback on SAA. In particular, students performing lower on a task usually use less valid cues for their self-assessment, as they are, for example, less familiar with the task and, therefore have less valid reference points and cues (Ehrlinger et al., 2008; Händel et al., 2020). Accordingly, lower-performing students might benefit more from receiving diagnostic cues (e.g., feedback) than higher-performing students, who tend to use diagnostic cues more frequently (Thiede et al., 2010). In line with this, few studies investigated a potential interaction between students' performance level and feedback effects and found that lower-performing students improved their SAA more with the help of feedback than higher-performing students did (Nederhand et al., 2019; Urban & Urban, 2021).

What has been largely overlooked is that the link between performance and SAA (i.e., lower-performing students are often lower-calibrated) can imply two distinct moderation effects: not only can performance levels moderate the effect of feedback, but so can initial SAA levels. Specifically, lower-calibrated students who usually perform worse also tend to use fewer diagnostic cues (Thiede et al., 2010) and, therefore might benefit more from receiving external cues through feedback. Such evidence would contribute to and extend the literature examining the moderation of performance levels (Nederhand et al., 2019; Panadero et al., 2016). In theoretical terms, it would emphasize that feedback communicates diagnostic cues and serves as an anchor for students' self-assessment (Koriat, 1997; Nicol, 2021), thereby highlighting that feedback can support those who are most in need of assistance.

1.4. The present study and research questions

Research is needed to determine whether the potential benefits of LLM-generated feedback also apply to SAA. LLM-generated feedback differs from human feedback in several ways—for instance, it is often written rather than numerical, may be less accurate (Steiss et al., 2024), or can include less valid grading scores (Liew & Tan, 2024). These characteristics may affect how students process the feedback and, in turn, its impact on SAA. For instance, evaluative feedback with performance scores seems to be particularly beneficial for SAA, as it provides clear benchmarks that help students to calibrate their self-assessments more accurately (Liebenow et al., 2024). Therefore, the present study investigated whether findings on feedback effects from previous research also hold for LLM-generated feedback by investigating the following research question:

RQ1. Does LLM-generated feedback enhance students' SAA?

To examine this research question, we conducted a randomized control experiment, including two experimental conditions. Students in both conditions wrote a first draft on a given writing task in English as a foreign language. After finishing their draft, they were instructed to revise their texts. Students in an experimental group received LLM-generated feedback during their text revision. Students in a control group did not receive feedback during their revision. Furthermore, we measured students' writing performance and self-assessment after finishing their first draft (pretest) and after completing their draft revision (posttest). To examine the effect of feedback, we tested the following hypothesis by analyzing differences in students' SAA changes between pretest and posttest:

H1. Students receiving LLM-generated feedback improve their SAA more compared to students who do not receive feedback.

To further enrich our understanding of feedback effects on SAA, it is crucial to investigate which students benefit most from feedback. Therefore, we consider potential moderators and investigated the following research question:

RQ2. Does feedback have differential effects depending on students' performance and SAA levels?

As outlined above, lower-performing students are usually less accurate in their self-assessment and are, therefore, in greater need of support (Ehrlinger et al., 2008; Thiede et al., 2010). But more importantly, lower-calibrated or lower-performing students might benefit more from additional diagnostic cues provided through feedback. Higher-performing students with more accurate self-assessment are already more likely to identify and utilize available diagnostic cues in the task or context itself (Händel et al., 2020; Thiede et al., 2010), which might limit the added value of external feedback. In contrast, lower-performing or less-calibrated students tend to overlook or misinterpret such cues and are, therefore, more reliant on external feedback to support their SAA. Following that, we investigated the following hypotheses:

H2a. Lower-performing students can improve their SAA more with LLM-generated feedback compared to higher-performing students.

H2b. Students with lower SAA improve their SAA more with LLM-generated feedback compared to students with higher SAA.

2. Methods

2.1. Transparency and openness

All supplements, data, and analysis scripts necessary to replicate our results can be found at: https://osf.io/tfrbg/?view_only=7dec2a0a6c674038b3c9d83cc20438ab. 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. The present study was not preregistered.

2.2. Study sample

This study reports secondary data analyses. Data were collected from May to June 2023, resulting in a sample of 545 10th-grade students. Participants were recruited from a selection of academic-track schools in the German federal state of Schleswig-Holstein. As an incentive for participation, schools were promised a report summarizing the aggregated results specific to their school. Due to technical issues, 86 students were excluded from the final analysis. These issues included participants not receiving feedback on their texts, survey crashes occurring before participants could compose their first text, and failures to save written texts. Consequently, the final data set included 459 students (52.75 % female, $M_{age} = 16.01$).

2.3. Study design and procedure

Students were randomly assigned to a control group (CG: $n = 256$) and an experimental group (EG: $n = 203$) within classrooms. The experiment was conducted on computers in an online survey format. After a short introduction, students in both groups worked on an argumentative writing task in English for a maximum of 20 min. After finishing their draft, all students self-assessed their writing performance (pretest). Students in the experimental group then received LLM-generated feedback on their draft, along with instructions to revise their text based on the feedback (i.e., "Please revise your text with the help of the feedback shown in the table below as best as you can. Take sufficient time for your revisions."). In the control group, which did not receive any feedback, students likewise had the opportunity to revise their text. However, instead of feedback, they were shown a standardized prompt: "This is the feedback for your text: Please read your argumentative essay again and try to revise the text as best you can. Take sufficient time for your revisions.". Importantly, control-group students were not given any actual feedback content; this generic instruction was displayed simply to prompt them to review and improve their drafts in the absence of feedback, thereby maintaining parity in the revision procedure between conditions. Furthermore, in both conditions, the actual task was displayed below the instruction, and the students' first drafts were shown in a window in which they could directly revise their texts. Additionally, students in both conditions had a maximum of 20 min to complete their revision. Therefore, the only differences between both conditions relate to the instruction for the revision and the feedback table in the experimental group. After completing their text revision, students self-assessed their revised writing performance (posttest). At the end of the experiment, we collected several demographic and background variables (i.e., age,

Aspect	Hints for improvement	Examples for improvement
structure	The essay lacks an introduction and a conclusion. The ideas presented are not clearly structured, making it difficult for the reader to follow. Use transition words to connect different ideas.	Include an introduction and a conclusion. Use paragraphs to group related ideas. Use transition words such as "however", "therefore", and "on the other hand" to connect different ideas.
content	The essay lacks specific examples to support the argument. The point of view is not clear. The essay is too general.	Include specific examples to support the argument. Make your point of view clear and support it throughout the essay. Be specific and avoid generalizations.
language	The essay contains many spelling and grammar errors. The vocabulary is limited. Sentences are too long and need to be broken down into smaller units.	Proofread the essay for spelling and grammar errors. Use a wide range of vocabulary. Break down long sentences into smaller units. Use connecting words such as "however", "in addition", and "nevertheless" to add clarity to the text.

Fig. 1. Feedback example generated by GPT

gender, parental education, number of books at home, and last English grade). The entire experiment was completed within a 90-min class period.

2.4. Materials

2.4.1. LLM feedback

In line with established feedback design principles (e.g., Hattie & Timperley, 2007; Narciss, 2006), we designed the prompt to provide feedback using the LLM GPT-3.5-turbo (which we refer to simply as GPT in the rest of this article). In detail, GPT was prompted to generate elaborate feedback, including hints and examples for text revision. These hints and examples were given separately for three text quality aspects: structure, content, and language. To reduce cognitive load, GPT was instructed to give this feedback in a tabular format. Furthermore, we made the context explicit, that is, that the English text was written by a foreign-language learner from upper secondary school, and, therefore, the feedback should address this context and recipient. The precise prompt provided to GPT can be found in the Appendix (Appendix A). GPT was embedded in the online survey via the application programming interface (API), with the following model settings: model: GPT-3.5-turbo, temperature: 0, maximum length: 1,800. A feedback example generated by GPT can be found in Fig. 1.

2.4.2. Writing task

Students were instructed to write an argumentative essay on one of two different tasks designed by the Test of English as a Foreign Language (TOEFL iBT®). In both tasks, students were asked to discuss a statement and to state, explain, and justify their own opinion. The statement for the first task 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." The statement for the second task 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." The two tasks were randomized between students and task order. Previous research has indicated that both tasks are similar in terms of their level of difficulty (Keller et al., 2020, Rupp et al., 2019).

2.5. Measures

2.5.1. Writing performance

All written texts were automatically scored by an algorithm. This algorithm was trained using training data containing 9,628 texts from 2,420 students (58.1 % female, $M_{age} = 17.7$ years). Expert raters, who underwent rigorous training, assessed each text using a comprehensive scale from zero (indicating low quality) to five (indicating high quality). The raters achieved exact agreement on 62.5 % of the texts and demonstrated a quadratic weighted kappa of 0.67. In cases where the raters disagreed, a third expert with extensive experience as a senior rater determined the final score. To train the algorithm, text quality was initially predicted on more than 100 linguistic features such as lexical information (e.g., word n-grams), structural information (e.g., part-of-speech n-grams), as well as features related to length and complexity. The text corpus is described in detail in Keller et al. (2020) and the rater training and the rating process are described in Rupp et al. (2019). A detailed description of the algorithm and the training process can be found in Meyer et al. (2024), Rupp et al. (2019), and Zesch and Horbach (2018). Following this automatic scoring, all students—including students in the control group—received summative feedback after the intervention. This approach adheres to ethical standards by maintaining the practice of giving all students balanced chances for educational growth and improvement.

2.5.2. Self-assessment accuracy

Students self-assessed their writing performance after writing the first draft (pretest) and after their revision (posttest) on a Likert scale ranging from zero (*very bad*) to six (*very good*) with the item "Please consider the text you have just written. How would you assess the quality of the text?" We measured SAA as the absolute difference between self-assessment and actual writing performance (i.e., lower scores indicated higher accuracy), following equation (1) from Schraw (2009). In addition to the absolute accuracy, we measured students' bias as the simple difference between their self-assessment and writing performance (i.e., positive scores indicate overestimation and negative scores indicate underestimation) by following equation (3) from Schraw (2009). We incorporate this measure to enhance the understanding of changes in students' SAA. By examining students' bias, we want to investigate whether the postulated improvement in students' SAA can be explained by a decreased under- or over-estimation in students' self-assessment. However, we want to note that bias scores do not reflect students' accuracy in absolute terms and can, therefore, be misleading in following conclusions (Liebenow et al., 2024; Schraw, 2009). Before calculating the SAA scores, we transformed students' self-assessment scores to align them with the measured writing performance scale, which ranged from 0 to 5. Specifically, the original self-assessment scores were divided by the maximum possible score (i.e., 6) and subsequently multiplied by 5.

2.5.3. Background variables

To control for confounding differences between students' demographics, we measured students' age, gender (1 = female, 0 = male), last English grade (lower scores indicating higher grades), parental education, and number of books at home as an indicator of socioeconomic status (Heppt et al., 2022).

2.6. Analytic approach

To investigate our research questions, we conducted hierarchical regression-based path analyses within a structural equation model (SEM) framework in R, Version 4.3.1 (R Core Team, 2023), using the R package *lavaan* (Rosseel, 2012) to consider missing data, by applying the full information maximum likelihood approach (FIML; see Enders, 2023). The extent of missing data varied from 0 % to 7.84 %. To test our first hypothesis (H1), we included in our testing models SAA changes or bias changes (i.e., differences between posttest and pretest scores) as criteria and the experimental conditions (EG = 1, CG = 0) as a predictor. When testing our second hypotheses (H2a and H2b), we included pretest writing performance, or pretest SAA, as an additional predictor (including the interaction term). In the vein of a robustness check, we conducted every analysis, including background variables (i.e., age, gender, English grade, parental education, and number of books at home) as additional predictors. Moreover, we conducted the same analyses using posttest SAA and bias as dependent variables while controlling for corresponding pretest scores. This allowed us to complete the picture and examine whether, beyond showing greater improvements in SAA and bias, students in the EG assessed themselves more accurately after receiving feedback (i.e., displaying smaller SAA and bias scores at posttest) than students in the CG.

3. Results

3.1. Descriptive statistics

The descriptive statistics revealed that, in line with our underlying moderation assumptions, students' SAA correlated with students' writing performance across both measuring points, indicating that lower-calibrated students showed less performance (see Table 1). Moreover, students' bias correlated negatively with performance, suggesting that higher overestimation relates to lower writing performance. Additionally, changes in SAA and bias correlated with changes in performance across both groups, indicating that students who improved their writing performance also improved their SAA and decreased their overestimation. In line with our assumptions relating to the feedback effect on SAA, we observed descriptively larger improvements in SAA and bias in the experimental condition, compared to the control condition (see Table 2). In terms of a manipulation check, we found no significant group differences for pretest SAA ($B = 0.33$, $SE = 0.31$, $p = .277$) and pretest bias ($B = 0.11$, $SE = 0.12$, $p = .365$). For more correlative and descriptive details with regards to the covariates, see Supplement A: https://osf.io/rm8xz.

3.2. Testing hypotheses

When conducting the hierarchical regressions to test our hypotheses, we found no significant main effect of feedback on changes in SAA (see Tables 3 and 4). Moreover, there was no significant main effect of feedback on students' change in bias (see Tables 5 and 6). Additionally, we conducted the same regression analyses with posttest SAA and posttest bias, and again, we found no evidence for a main effect of feedback on any SAA outcome (for a complete overview of our analyses, see Supplement: https://osf.io/rm8xz). Therefore, our findings did not support our first hypothesis (i.e., H1).

Next, we investigated our second research question and examined potential moderation effects between feedback and pretest SAA on SAA changes. In line with our hypothesis (H2a), we found a significant positive interaction effect (with and without covariates), indicating that students with lower pretest SAA improved their SAA more with feedback than students with similar pretest SAA and without feedback (see Table 3). Additionally, we found a significant negative interaction ($\beta = -0.29$, $SE = 0.07$, $p < .001$) between feedback and pretest SAA on *posttest* SAA, indicating that students with lower pretest SAA were more accurate with feedback than similar students without feedback (see Fig. 2a). Moreover, we found a similar pattern of results for bias, indicating that students improved their SAA by decreasing their initial overestimation (see Table 5 and Fig. 2b). Contrary to our H2b, we could not find a significant interaction effect between students' pretest writing performance and feedback on their SAA changes (see Table 4) and bias changes (see Table 6). This was also evident for posttest SAA and bias (see Supplement: https://osf.io/rm8xz). These findings did not change when both interaction effects were investigated simultaneously (see Supplement: https://osf.io/rm8xz).

Table 1
Means, standard deviations, and partial correlations.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Pre WP	2.45	0.96								
2. Post WP	2.78	0.90	0.65**							
3. WP Ch.	0.32	0.78	−0.48**	0.36**						
4. Pre SAA	1.67	3.14	−0.30**	−0.14**	0.21**					
5. Post SAA	1.47	2.47	−0.12*	−0.24**	−0.14**	0.48**				
6. SAA Ch.	0.20	2.93	−0.23**	0.04	0.33**	0.67**	−0.33**			
7. Pre Bias	0.06	1.29	−0.57**	−0.30**	0.35**	0.41**	0.21**	0.27**		
8. Post Bias	0.04	1.21	−0.29**	−0.46**	−0.18**	0.22**	0.25**	0.03	0.58**	
9. Bias Ch.	0.02	1.14	−0.34**	0.14**	0.58**	0.24**	−0.02	0.27**	0.51**	−0.40**

Note. WP = writing performance. Ch. = change. Smaller values of SAA indicate higher accuracy. Positive values in SAA change indicate improved SAA. Positive values in bias indicate overestimation. Positive values in bias change indicate a reduction in overestimation.
* indicates $p < .05$. ** indicates $p < .01$.

Table 2
Means and standard deviations of variables per group.

Variable	CG (n = 256)		EG (n = 203)	
	M	*SD*	*M*	*SD*
Pre WP	2.45	0.95	2.47	0.97
Post WP	2.70	0.88	2.87	0.92
Pre SAA	1.53	2.81	1.86	3.52
Post SAA	1.51	2.73	1.42	2.07
SAA Ch.	0.01	2.65	0.44	3.24
Pre Bias	0.01	1.24	0.12	1.36
Post Bias	0.07	1.23	0.01	1.19
Bias Ch.	−0.06	1.05	0.12	1.26

Note. WP = writing performance. Ch. = change. Note that smaller values of SAA indicate higher accuracy.

Table 3
Results of Hierarchical Regressions with SAA Changes as Criterion and pretest SAA as Moderator.

Predictor	Model 1	Model 2	Model 3	Model 4
EG > CG	0.07 (0.05)	−0.04 (0.04)	0.08 (0.05)	−0.03 (0.04)
Pre SAA		0.51*** (0.05)		0.61*** (0.05)
Interaction		0.24*** (0.06)		0.19*** (0.05)
Age			−0.05 (0.05)	−0.20*** (0.04)
Gender			−0.01 (0.05)	0.07* (0.03)
Books			0.08 (0.05)	0.09** (0.03)
Grade			−0.00 (0.05)	0.07* (0.04)
Mother Ed.			−0.07 (0.06)	−0.01 (0.04)
Father Ed.			0.18** (0.06)	0.04 (0.04)

Note. *b*, *SE* in parentheses relate to the standardized regression coefficients. SAA = SAA. Interaction = pretest SAA:group dummy variable (EG = 1, CG = 0). Books = number of books at home. Grade = last English grade. Ed. = education.
* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Table 4
Results of hierarchical regressions with SAA changes as criterion pretest writing performance as moderator.

Predictor	Model 1	Model 2	Model 3	Model 4
EG > CG	0.07 (0.05)	0.21 (0.13)	0.08 (0.05)	0.18 (0.13)
Pre WP		−0.18** (0.06)		−0.22*** (0.07)
Interaction		−0.15 (0.14)		−0.12 (0.14)
Age			−0.05 (0.05)	−0.06 (0.05)
Gender			−0.01 (0.05)	0.03 (0.05)
Books			0.08 (0.05)	0.09 (0.05)
Grade			−0.00 (0.05)	−0.05 (0.05)
Mother Ed.			−0.07 (0.06)	−0.07 (0.06)
Father Ed.			0.18** (0.06)	0.13* (0.06)

Note. *b*, *SE* in parentheses relate to the standardized regression coefficients. WP = writing performance. Interaction = pretest WP:group dummy variable (EG = 1, CG = 0). Books = number of books at home. Grade = last English grade. Ed. = education.
* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Table 5
Results of hierarchical regressions with bias changes as criterion and pretest bias as moderator.

Predictor	Model 1	Model 2	Model 3	Model 4
EG > CG	0.08 (0.05)	0.05 (0.04)	0.08 (0.05)	0.04 (0.04)
Pre Bias		0.41*** (0.05)		0.44*** (0.06)
Interaction		0.15* (0.06)		0.14* (0.06)
Age			0.02 (0.05)	−0.01 (0.04)
Gender			−0.01 (0.05)	0.08 (0.04)
Books			0.03 (0.05)	0.04 (0.04)
Grade			−0.01 (0.05)	0.05 (0.04)
Mother Ed.			0.01 (0.06)	0.00 (0.05)
Father Ed.			0.02 (0.06)	−0.05 (0.05)

Note. *b*, *SE* in parentheses relate to the standardized regression coefficients. Interaction = pretest bias:group dummy variable (EG = 1, CG = 0). Books = number of books at home. Grade = last English grade. Ed. = education.
* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Table 6
Results of hierarchical regressions with bias changes as criterion and pretest writing performance as moderator.

Predictor	Model 1	Model 2	Model 3	Model 4
EG > CG	0.08 (0.05)	0.26* (0.13)	0.08 (0.05)	0.25 (0.13)
Pre WP		−0.28*** (0.06)		−0.33*** (0.06)
Interaction		−0.19 (0.13)		−0.19 (0.13)
Age			0.02 (0.05)	0.00 (0.05)
Gender			−0.01 (0.05)	0.04 (0.05)
Books			0.03 (0.05)	0.05 (0.05)
Grade			−0.01 (0.05)	−0.09 (0.05)
Mother Ed.			0.01 (0.06)	0.01 (0.06)
Father Ed.			0.02 (0.06)	−0.04 (0.06)

Note. *b*, *SE* in parentheses relate to the standardized regression coefficients. WP = writing performance. Interaction = pretest WP:group dummy variable (EG = 1, CG = 0). Books = number of books at home. Grade = last English grade. Ed. = education.
* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

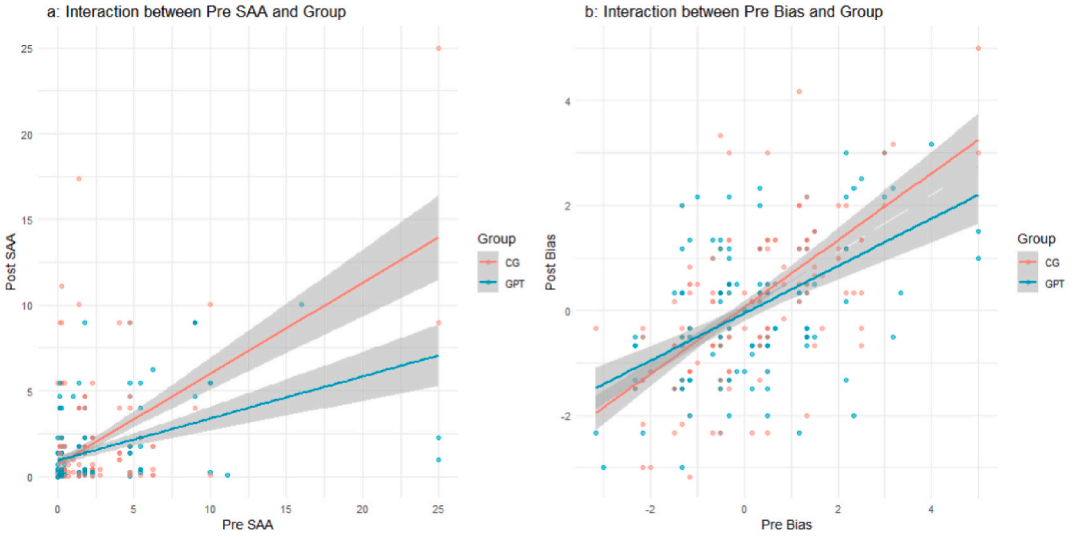

Fig. 2. Interaction Effects
Note. Smaller values of SAA indicatie higher accuracy. Negative values of bias indicate underestimation, and positive values indicate overestimation.

3.3. Exploratory analyses

As we found no significant main effects of feedback on any SAA measure but significant interaction effects between students' pretest SAA and bias, we further explored the interaction effects. By conducting simple slope analyses, we found a significant negative effect of feedback on changes in SAA ($B = -0.66$, $SE = 0.29$, $p = .03$) for students with higher pretest SAA (-1 $SD = -1.48$), indicating that these students decreased their SAA with feedback more, compared to students with similar higher pretest SAA and without feedback. Additionally, we conducted the same analysis with bias changes as criterion and pretest bias as moderator and found no significant effect of feedback on bias change for students with lower pretest bias (-1 SD = −1.23). Therefore, we cannot conclude whether students with higher pretest SAA decreased their accuracy due to an increase in over- or underestimation. However, we want to note that the lower score of pretest SAA (−1.48) estimated within the simple slope analysis was outside the range of possible and observed scores. Therefore, this result should be interpreted with caution and viewed as a tendency finding.

Moreover, we examined whether students with lower pretest SAA improved this accuracy through feedback—either because they decreased their initial overestimation or because their writing performance improved. To clarify this, we ran a regression analysis testing the interaction effect when controlling for changes in writing performance. This analysis revealed a significant interaction effect ($\beta = 0.24$, $SE = 0.05$, $p < .001$), suggesting that students with lower pretest accuracy significantly improved their SAA with feedback when having the same changes in writing performance.

As the interaction effects between pretest writing performance and feedback on SAA were not significant, we applied the Johnson-Neyman method (Johnson & Neyman, 1936) to further probe the interactions. Our post-hoc analyses revealed that the interaction between pretest writing performance ($M = 2.45$, $SD = 0.96$, range from 0 to 4) and feedback group was significant for students with pretest writing performance scores ranging from −0.14 to 2.37. Within this range, students improved their SAA significantly more with the help of feedback than similar students without feedback.

4. Discussion

Our study aimed to investigate the impact of LLM-generated feedback on SAA among upper-secondary students in the context of a writing task. To the best of our knowledge, the present study is the first randomized control experiment testing LLM-generated feedback for this purpose. In the next sections, we discuss our findings following our guiding research questions.

RQ1. Does LLM-generated feedback enhance SAA?

When investigating our first research question, we found no significant main effect of LLM-generated feedback on SAA or bias. In other words, on average, students who received feedback could not significantly improve their SAA or reduce their bias through feedback. This result may seem surprising given prior evidence that feedback can foster SAA (e.g., Liebenow et al., 2024).

One reason that might explain the missing effect of LLM-generated feedback, on average, could be the nature of the feedback content. The feedback provided was rich in specific comments on the essay (i.e., on argument structure, content, and language), but it lacked an explicit holistic score or evaluative benchmark. Students were not given an explicit indication of their overall performance level, which made it difficult for them to gauge "how am I going" from the feedback alone. In the absence of a concrete overall appraisal, our feedback may had limited potential to recalibrate students' self-assessment, reflecting a conservative estimate of the feedback effect. Research indicates that students use feedback as a cue to adjust their judgments, especially when it provides clear standards or norms for performance (Andrade et al., 2010; Panadero et al., 2016). Moreover, meta-analytic evidence indicates that performance feedback that communicates the knowledge of the result has larger effects than other types of feedback without this information (Liebenow et al., 2024). However, we decided not to prompt GPT to provide holistic performance scores within the feedback, as previous studies indicated that GPT-3.5 Turbo has limited grading capabilities (Liew & Tan, 2024). Nonetheless, more recent LLMs have shown promising results in this regard (Yavuz et al., 2025), reinforcing our interpretation that our findings reflect a conservative estimate of LLM-generated feedback effectiveness.

Moreover, many students likely focused on using the comments to improve their essays (a task-oriented use of feedback) rather than for reflection and recalibrating their self-assessment. As scholars argue, when feedback predominantly targets the task level without addressing or stimulating self-assessment processes, effects on such processes become unlikely (Kluger & DeNisi, 1996; Stone, 2000). In line with that, studies found that feedback focusing on students' self-assessment is more beneficial for SAA (Urban & Urban, 2021; Van Loon & Roebers, 2020). Thus, the design of the feedback might have limited its average effect on SAA. However, our results are consistent with studies demonstrating that external feedback does not invariably improve self-assessment processes (Panadero et al., 2023). In some cases, feedback interventions have yielded minimal gains in accuracy or even led to lower subsequent SAA (Ernst et al., 2025; Raaijmakers et al., 2019). Thus, simply providing LLM-generated feedback was not a universally effective strategy for improving students' SAA. This lack of an overall effect suggested that the influence of feedback might depend on certain factors, prompting our second research question.

RQ2. Do lower-performing or less-calibrated students benefit more from feedback?

In line with our assumptions, the effectiveness of the LLM-generated feedback depended on students' levels of SAA. We observed significant interaction effects indicating that students with lower pretest SAA did improve their SAA more (i.e., SAA change) after receiving feedback and were more accurate afterward (posttest SAA) than similar students without feedback. Moreover, students who initially overestimated their performance (i.e., pretest bias) exhibited a marked reduction in their overestimation (i.e., bias change) and an increased accuracy (i.e., smaller posttest bias) after receiving feedback. However, the question arises, whether lower-calibrated students became more accurate, as they improved their self-assessment or were less overestimated because they increased their performance (Kornell & Bjork, 2009; Raaijmakers et al., 2019). To address this significant concern, which has rarely been addressed in the literature, we further examined whether the observed interaction effect persists when controlling for performance gains. Following our results, lower-calibrated students enhanced their SAA with feedback and not only their performance, as the interaction effect remains significant when looking at students with the same extent of performance changes. Accordingly, this interaction pattern emerged consistently across different outcome measures and analysis approaches. This robustness strengthens our confidence that students' pretest SAA moderated the effect of LLM-generated feedback on SAA. Therefore, our findings support the idea that feedback acts as a more powerful catalyst for those who have a greater need for support.

Students with lower initial SAA typically do not use diagnostic cues to self-assess their performance and thus stand to gain the most from external input, as more accurate students tend to use already diagnostic cues (Thiede et al., 2010). Kruger and Dunning (1999) pointed out that less accurate students are also less aware of their miscalibration (i.e., "the double curse of incompetence"), indicating that especially those would profit from external cues and feedback that makes them more aware of their miscalibration. From the cue-utilization perspective (Koriat, 1997), our feedback likely supplied diagnostic cues that lower-accurate students could finally recognize and use to adjust and correct their self-assessment. In detail, the LLM-generated feedback provided individualized comments on written essays, which may have served as an anchoring reference point for those who were less calibrated. Consistent with the anchor effect of feedback (Panadero et al., 2023; Nicol, 2021), students with lower pretest SAA seem to have adjusted their self-assessment in light of the feedback information that prompted students to notice issues they initially were unaware of.

Contrary to our hypotheses, our analyses did not reveal a significant interaction effect between pretest writing performance and feedback on any SAA measures (i.e., SAA changes and posttest SAA). However, explorative simple slope analyses revealed that lower-performing students who received feedback improved their SAA and bias changes more strongly than similar students without feedback. This effect did not occur among higher-performing students, indicating that feedback did not affect the SAA of higher-performing students. Further Johnson-Neyman analyses confirmed this pattern: group differences became significant primarily at

lower performance levels.

These findings align with previous studies. A recent study conducted by Theobald and Brod (2025) found no significant interaction effect between performance levels and feedback on SAA and attributed this to insufficient statistical power—a consideration that likely applies to our study as well. Similarly, Urban and Urban (2021) found no significant interaction effect between feedback and performance level on SAA. However, they reported that lower-performing students benefited more from feedback than their peers without feedback. It is important to note that interaction effects in educational research are often small or non-significant, especially in regression models due to measurement errors that are amplified in the interaction term (Busemeyer & Jones, 1983; Nagengast et al., 2011). All in all, our regression analyses did not support our hypotheses regarding the moderating role of initial performance levels. Only posthoc analyses suggested that lower-performing students may still benefit more from feedback.

As feedback was not effective on average (i.e., no significant main effects of feedback), a question one could ask is, whether students with higher initial SAA might derived negligible benefit from the feedback or even might become less accurate in their self-assessment following feedback. Therefore, we examined this question within the exploratory analysis and found evidence that students with higher pretest SAA (-1 SD) significantly reduced their SAA (i.e., showing less posttest and changes SAA) with feedback than similar students without feedback. This pattern of results might explain why our analyses revealed no significant effect of feedback on average. However, this result should be interpreted cautiously, as the estimated score for a lower pretest SAA score (−1.48) lies outside the possible score range.

Higher-performing or well-calibrated students have less room for improvement; since their initial self-assessments were already close to their actual performance, it is unsurprising that feedback did not substantially boost their accuracy further (Nederhand et al., 2019). What is surprising is that our feedback might have interfered with or disrupted their accurate self-assessment. One potential explanation lies within our specific feedback. We prompted the LLM to provide feedback regarding potential points for improvement on the given writing criteria without explicitly stating which aspects have already been fulfilled. Therefore, our feedback was purely formative – it offered comments on the essay's content and form but no performance score or comparison to students' self-assessment. For students with higher pretest SAA and performance, such feedback might highlight minor flaws without affirming the overall high quality of the work. This could lead such students to focus on the critique and, therefore, underestimate their performance after receiving feedback. In fact, there is evidence that receiving feedback can sometimes narrow students' focus on the feedback content in a way that hampers their broader self-assessment process. Panadero et al. (2020, 2023) found that students who received feedback subsequently used fewer self-assessment strategies and criteria and tended to focus only on aspects mentioned in the feedback. In our context, higher-accurate students may have concentrated on the LLM's comments (e.g., a suggestion to improve a certain paragraph) and neglected their own holistic self-assessment, resulting in a decline in SAA. This interpretation underscores the potential importance of how feedback is delivered, especially when considering individual differences.

4.1. Summary

In summary, our study provides nuanced empirical evidence for the role of LLM-generated feedback in student self-assessment. Rather than uniformly boosting SAA for all students, the feedback effect was revealed when individual differences were considered. The LLM-generated feedback operated as a targeted aid, substantially improving the SAA of students who initially overestimated their performance. Our results, therefore, partially contributed to theoretical frameworks, emphasizing that feedback can serve as an effective mechanism that helps students to identify gaps between their self-assessed performance and actual performance outcome (Butler & Winne, 1995; Panadero et al., 2024). Moreover, these findings reinforce the idea that students with more need for support benefit the most from feedback with regard to SAA (Nederhand et al., 2019). Therefore, our results extend the growing body of evidence into the realm of LLM-generated feedback that considers individual differences when investigating feedback effects on self-assessment (e.g., Liebenow et al., 2024; Panadero et al., 2023; Theobald & Brod, 2025). They demonstrate that even though feedback is delivered automatically by an LLM, it can effectively provide individualized support in real-time to those who struggle with self-assessment. At the same time, feedback did not boost (and in some cases impeded) the SAA of students who were already relatively accurate. This emphasizes the need for carefully designed adaptive feedback within practical implementation rather than one-size-fits-all solutions.

4.2. Limitations and implications

One major limitation of our study could be attributed to our prompted feedback design. Although our feedback design was based on empirical evidence and theory (e.g., Hattie & Timperley, 2007; Mertens et al., 2022), our study provided only one type of feedback. As discussed above, we provided elaborated feedback without any explicit information that could be classified as Knowledge of Result feedback [KR] (Narciss, 2006), which might have harmed students' SAA who initially self-assessed their performance more accurately. It is, therefore, crucial that future studies delve deeper into the specific design features that might determine the effectiveness of feedback on students' SAA. As recommended by Liebenow et al. (2024), future research needs to investigate and compare the effects of different types of feedback or their combinations to identify which types or combination is most beneficial to foster students' SAA. Moreover, our feedback did not communicate any information in relation to students' prior SAA that may have limited the observed effects, as self-assessment-focused feedback can be favorably compared to performance-focused feedback (Panadero et al., 2016; Urban & Urban, 2018).

One notable limitation is the brevity of the intervention, which was delivered in a single 90-min session. While this short timeframe minimized exposure to external confounding variables, it constrained opportunities for extended self-assessment processes and

reflection beyond the immediate session. Nevertheless, within this limited duration, lower-calibrated students improved their SAA with feedback, suggesting that even brief feedback can engender a feedback-driven recalibration of their self-assessment. This pattern aligns with observations by Correnti et al. (2022), who found that students' post-feedback reflections often mirrored the language and content of the feedback they received, indicating learning progreess through feedback. However, future research should investigate whether these gains in SAA persist over time and explore in greater depth how students process and internalize feedback across longer periods and multiple feedback cycles.

Another limitation is the "weak" control group, as students in the control group did not receive any feedback for their text revision. It is thus difficult to fully assess the relative effectiveness of the LLM-generated feedback. Therefore, future research could compare our LLM-generated feedback with similar feedback from other agents (e.g., teachers or automated systems) to assess the extent to which our findings are attributable to characteristics specific to LLMs, which inherently come with particular strengths and limitations. For instance, LLM-generated feedback has the benefit over traditional automated feedback in that it can be delivered independent of tasks and learning domains (Fleckenstein et al., 2024), which allows for the quick and easy implementation of feedback. On the downside, LLM-generated feedback carries the risk of communicating hallucinated false content (Farquhar et al., 2024), which—in turn—could negatively influence the effect on SAA. Nonetheless, our control group can be considered to be ecologically valid, as the provision of real-time feedback for revision is typically not feasible in the context of writing due to the constraints of limited resources and time.

The implications of our study apply specifically to GPT-3-turbo and cannot be generalized to other LLMs. Given rapid advances in the field, it is essential to compare different models to assess whether our findings are model-specific (Chiu et al., 2023). Moreover, future research should compare previous versions with current versions of LLMs to predict the performance of upcoming models. Notably, GPT-3-turbo is already outdated at the time of writing, making our results a conservative estimate of what newer models may achieve in terms of feedback and SAA.

4.3. Conclusion

The present study enriches the literature by offering empirical evidence on the potential advantages of using LLMs in educational settings (Kasneci et al., 2023; Yan et al., 2024). In particular, our study provided nuanced evidence on the effectiveness of LLM-generated feedback in relation to students' SAA. These findings contribute to research investigating students' performance levels as a potential moderator on the effect of feedback on SAA (e.g., Liebenow et al., 2024; Nederhand et al., 2019). Moreover, our study offers a new potential moderator to the current research discourse - students' SAA. It suggests that for lower-calibrated students who need the most support, LLM-generated feedback offers a cost-effective and scalable means to enhance SAA. This can be interpreted as a Robin-Hood effect, where students who are more in need of support profit more from an intervention than higher-performing (e.g., higher-calibrated) students (Häfner et al., 2017).

At the same time, we found that feedback design and implementation should be carefully considered. Feedback may be counterproductive if it focuses solely on areas for improvement without answering the question of "How am I going?" (Hattie & Timperley, 2007). Accordingly, we recommend a careful implementation of LLM-generated feedback in classrooms by incorporating these insights to empower students in their SAA. While such feedback has the potential to save significant time and resources, further research is needed to better understand how different feedback types affect students' SAA, particularly when generated by LLMs.

CRediT authorship contribution statement

Lucas W. Liebenow: Writing – original draft, Visualization, Software, Formal analysis, Data curation, Conceptualization. **Fabian T.C. Schmidt:** Writing – review & editing, Supervision. **Jennifer Meyer:** Writing – review & editing, Project administration. **Johanna Fleckenstein:** Writing – review & editing, Project administration, Funding acquisition.

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

During the preparation of this work, the first author used GPT-4-omni, designed by OpenAI, to improve language and readability. After using this tool, the first author reviewed and edited the content as needed. The first author takes full responsibility for the content of the publication.

Funding sources

This work was supported by the junior research group "Formative Writing Assessment: Automated Feedback Using Artificial Intelligence" (FORMAT) and funded by the Federal Ministry of Education and Research (grant number: 01JG2104).

Declaration of competing interests

The authors have no conflict of interest to disclose.

Acknowledgements

We would like to thank Gráinne Newcombe for her assistance with language editing. Jennifer Meyer is supported by a Jacobs

Foundation Research Fellowship (2024–2026).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2025.105385.

Data availability

All supplements, data, and analysis scripts necessary to replicate our results can be found at: https://osf.io/tfrbg/?view_only=7dec2a0a6c674038b3c9d83cc20438ab. 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. The present study was not preregistered.

References

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17*(2), 199–214. https://doi.org/10.1080/09695941003696172

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*(1), 133–144. https://doi.org/10.1016/0749-5978(87)90049-5

Baas, D., Castelijns, J., Vermeulen, M., Martens, R., & Segers, M. (2015). The relation between assessment for learning and elementary students' cognitive and metacognitive strategy use. *British Journal of Educational Psychology, 85*(1), 33–46. https://doi.org/10.1111/bjep.12058

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290.

Braumann, S., van de Pol, J., Kok, E., Pijeira-Díaz, H. J., van Wermeskerken, M., de Bruin, A. B. H., & van Gog, T. (2024). The role of feedback on students' diagramming: Effects on monitoring accuracy and text comprehension. *Contemporary Educational Psychology, 76*, Article 102251. https://doi.org/10.1016/j.cedpsych.2023.102251

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1–13. https://doi.org/10.1016/j.iheduc.2015.04.007

Brown, G., & Harris, L. (2014). The future of self-assessment in classroom practice: Reframing self- assessment as a core competency. *Frontline Learning Research*. https://doi.org/10.14786/flr.v2i1.24

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93*(3), 549–562. https://doi.org/10.1037/0033-2909.93.3.549

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open, 3*, Article 100084. https://doi.org/10.1016/j.caeo.2022.100084

De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

De Bruin, A. B. H., & Van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1–9. https://doi.org/10.1016/j.learninstruc.2017.06.001

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. https://doi.org/10.1037/met0000563

Ernst, H. M., Prinz-Weiß, A., Wittwer, J., & Voss, T. (2025). Discrepancy between performance and feedback affects mathematics student teachers' self-efficacy but not their self-assessment accuracy. *Frontiers in Psychology, 15*, Article 1391093. https://doi.org/10.3389/fpsyg.2024.1391093

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. https://doi.org/10.3102/00346543059004395

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

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*, Article 1162454. https://doi.org/10.3389/frai.2023.1162454

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6*, Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (677, Vol. 1, p. 647). Cambridge University Press. https://doi.org/10.1017/9781108235631.026.

Häfner, I., Flunger, B., Dicke, A.-L., Gaspard, H., Brisson, B. M., Nagengast, B., & Trautwein, U. (2017). Robin hood effects on motivation in math: Family interest moderates the effects of relevance interventions. *Developmental Psychology, 53*(8), 1522–1539. https://doi.org/10.1037/dev0000337

Händel, M., De Bruin, A. B. H., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning, 15*(1), 51–75. https://doi.org/10.1007/s11409-020-09220-0

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

Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education, 25*(4), 903–928. https://doi.org/10.1007/s11218-022-09704-8

Johnson, P., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. https://www.semanticscholar.org/paper/Tests-of-certain-linear-hypotheses-and-their-to-Johnson-Neyman/a89c0d064ac8cf681ebfadb7e53ad45403c2c31d.

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. https://doi.org/10.35542/osf.io/5er8f.

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*, 100700. https://doi.org/10.1016/j.jslw.2019.100700

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In *Handbook of metamemory and memory* (pp. 117–135). Psychology Press.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449–468. https://doi.org/10.1037/a0017350

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), 173–194. https://doi.org/10.1007/s11409-010-9056-2

León, S. P., Panadero, E., & García-Martínez, I. (2023). How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educational Psychology Review, 35*(4), 106. https://doi.org/10.1007/s10648-023-09819-0

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior & Human Performance, 26*(2), 149–171.

Liebenow, L. W., Schmidt, F. T. C., Meyer, J., Panadero, E., & Fleckenstein, J. (2024). Supporting self-assessment: A systematic review and meta-analysis of feedback effects. https://doi.org/10.31219/osf.io/k94eq.

Liew, P. Y., & Tan, I. K. T. (2024). On automated essay grading using large language models. *Proceedings of the 2024 8th international conference on computer science and artificial intelligence* (pp. 204–211). https://doi.org/10.1145/3709026.3709030

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., 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.* , Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "×" out of expectancy-value theory?: A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science, 22*(8), 1058–1066. https://doi.org/10.1177/0956797611415540

Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse.* Waxmann.

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. J. P. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology, 33*(6), 1068–1079. https://doi.org/10.1002/acp.3548

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5. Elsevier.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*(2), 159–179. https://doi.org/10.1007/s10409-006-9595-6

Nilsson, F., & Tuvstedt, J. (2023). GPT-4 as an automatic grader: The accuracy of grades set by GPT-4 on introductory programming assignments. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-330993.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Alonso-Tapia, J., & Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies In Educational Evaluation, 39*(3), 125–132.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education, 44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review, 28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., Fernández, J., Pinedo, L., Sánchez, I., & García-Pérez, D. (2024). A self-feedback model (SEFEMO): Secondary and higher education students' self-assessment profiles. *Assessment in Education: Principles, Policy & Practice*, 1–33. https://doi.org/10.1080/0969594X.2024.2367027

Panadero, E., Fernández-Ruiz, J., & Sánchez-Iglesias, I. (2020). Secondary education students' self-assessment: The effects of feedback, subject matter, year level, and gender. *Assessment in Education: Principles, Policy & Practice, 27*(6), 607–634. https://doi.org/10.1080/0969594X.2020.1835823

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E., Pérez, D. G., Ruiz, J. F., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. L. (2023). University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education, 38*(3), 1031–1051. https://doi.org/10.1007/s10212-022-00639-4

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning, 14*(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

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. https://doi.org/10.1002/ets2.12249

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Schwam, D., Greenberg, D., & Li, H. (2021). Individual differences in self-regulated learning of college students enrolled in online college courses. *American Journal of Distance Education, 35*(2), 133–151. https://doi.org/10.1080/08923647.2020.1829255

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

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. https://doi.org/10.1023/A:1009084430926

Theobald, M., & Brod, G. (2025). The role of feedback and working memory for goal-related monitoring and goal revision. *Learning and Instruction, 97*, Article 102108. https://doi.org/10.1016/j.learninstruc.2025.102108

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362.

Urban, K., & Urban, M. (2018). Influence of fluid intelligence on accuracy of metacognitive monitoring in preschool children fades with the calibration feedback. *Studia Psychologica, 60*(2), 123–136. https://doi.org/10.21909/sp.2018.02.757

Urban, K., & Urban, M. (2021). Effects of performance feedback and repeated experience on self-evaluation accuracy in high- and low-performing preschool children. *European Journal of Psychology of Education, 36*(1), 109–124. https://doi.org/10.1007/s10212-019-00460-6

Van Loon, M. H., & Roebers, C. M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. *Early Childhood Research Quarterly, 53*, 301–313. https://doi.org/10.1016/j.ecresq.2020.05.007

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*(10), 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*(1), 150–166. https://doi.org/10.1111/bjet.13494

Zesch, T., & Horbach, A. (2018). ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, … T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1365/.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. https://doi.org/10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7.