Alemdag and Narciss
Int J Educ Technol High Educ (2025) 22:22
https://doi.org/10.1186/s41239-025-00522-4

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Promoting formative self-assessment through peer assessment: peer work quality matters for writing performance and internal feedback generation

Ecenaz Alemdag[1,2*] and Susanne Narciss[1,3]

*Correspondence:
ecenazalemdag@gmail.com

[1] Psychology of Learning and Instruction, Technische Universität Dresden, Dresden, Germany
[2] Institute of Educational Sciences, Europa-Universität Flensburg, Flensburg, Germany
[3] Center of Tactile Internet with Human in the Loop (Ceti), Technische Universität Dresden, Dresden, Germany

Abstract

Generating quality and accurate internal feedback in self-assessment is challenging for higher education students. Assessing peer works has been linked to enhancing concurrent and follow-up self-assessments in prior research and theoretical models. However, empirical evidence concerning how peer work quality influences subsequent behaviors in online learning environments is scarce and contradictory. Hence, this study aimed to investigate the effects of the quality of the assessed peer works on writing performance and the amount, content, and accuracy of internal feedback via an online experiment. The participants were 164 university students who completed abstract writing and assessment tasks in an online setting. Specifically, they wrote an abstract for a study; provided feedback on two low-quality peer abstracts (PF− −), one high-quality and one low-quality (PF+ −) abstract, or two high-quality abstracts (PF+ +); self-scored and generated internal feedback for their abstracts; and revised their abstracts. The groups' initial and revised abstracts, self-scorings, and internal feedback were analyzed. The findings revealed that the PF+ + and PF+ − groups had significantly higher revised writing performance, more feedback comments expressing problems in their writings and proposing specific suggestions for improvement, and more accuracy in self-scoring and problem detection than the PF− − group. Indicating the impact of the peer work quality, this study suggests using at least one high-quality work for (online) peer assessment to enhance writing performance and self-assessment accuracy.

Keywords: Feedback, Self-assessment, Peer assessment, Formative assessment, Exemplars, Writing, Online learning

Introduction

Formative self-assessment is a process during which students reflect on the quality of their performance, evaluate to what extent their performance fulfills expected standards, and revise them based on their judgments (Andrade & Valtcheva, 2009). Several systematic reviews and meta-analyses (e.g., Andrade, 2019; Brown et al., 2013; Panadero et al., 2017; Yan et al., 2023) reported significant contributions of self-assessment

Springer Open

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

to students' achievement, self-regulated learning skills, metacognition, and self-efficacy beliefs at different educational levels. However, these benefits occur especially when students generate quality and accurate internal feedback that enhances the learning process and performance (Andrade, 2019). Nonetheless, it might not be possible for all learners, particularly those with limited knowledge about assessment criteria and standards and low task expertise (Narciss, 2017; Kostons et al., 2009; Panadero et al., 2016, 2019). Moreover, the latest meta-analysis by León et al. (2023) revealed lower self-assessment accuracy in higher education compared to primary and secondary education. Therefore, it is necessary to support students, especially in higher education, in generating quality and accurate internal feedback (Peters et al., 2018; Panadero et al., 2016).

Assessing and generating feedback for peer works is one strategy that can complement self-assessment, but the issue of how the quality of peer works influences internal feedback generation and performance improvement is under-investigated (Nicol & McCallum, 2022), and there are contradictory findings (e.g., Cho & Cho, 2011; Dmoshinskaia et al., 2022; Wu et al., 2022). Strong empirical findings on this issue can inform decisions on how to allocate different works among students for peer assessment to optimize learning benefits. It can be helpful notably for the assessment design of (massive open) online courses that include a high number of students and rely on self- and peer assessment for grading and increasing student engagement and learning. Addressing this gap, the current study examined the effects of the quality of the assessed peer works on writing performance and internal feedback in an online learning environment. The amount, content, and accuracy of internal feedback were analyzed to provide a comprehensive account of the influence of the intervention on self-assessment.

Literature

Self-assessment and internal feedback

Self-assessment refers to a process during which students gather information concerning their learning process and outcomes and evaluate their quality based on specific criteria (Yan & Brown, 2017). It can result in self-scoring, numerical values assigned by learners themselves, and qualitative information that can describe their strengths and weaknesses (Panadero et al., 2019). Both kinds of information can be called self-feedback or internal feedback learners generate while comparing their current competence, processes, and products with reference information or values (Narciss, 2017; Nicol, 2021). However, since the latter can carry more explicit and rich information to enhance the learning process and performance, it might better serve the learning-oriented and formative purpose of self-assessment (Andrade, 2019).

Aligned with this purpose, internal feedback has been placed at the core of the regulation of learning since Butler and Winne's (1995) prominent work. The recent model by Nicol (2021) also explains internal feedback generation in self-regulated and co-regulated learning and highlights the importance of comparison in this process. According to this model, first, learners define their goals and plans for a task and then determine their tactics and strategies to generate internal and external products. While developing these products, they generate internal feedback by comparing their products with their goals. However, internal feedback also results from using external information for comparison. External information for this process includes rubrics, exemplars, peer works,

peer comments, and class notes. In the self-feedback model by Panadero et al. (2024), assessing and comparing with these kinds of information have also been indicated as one of the main phases of self-assessment performance based on empirical data. The internal feedback generated from comparisons can finally lead students to redo the tasks, modify their knowledge and beliefs, and enhance their performance. This study used peer works in peer assessment as exemplars to enhance internal feedback generation for writing tasks.

Writing and internal feedback

Formal activities for internal feedback generation can be incorporated into various subject domains and tasks in different educational settings. However, writing is one of the most prevalent works used in research on self-assessment (Andrade & Valtcheva, 2009). Likewise, this study used an academic writing task for self-assessment. Internal feedback generation in writing can support learners' writing development, metacognition, autonomy, self-efficacy, classroom collaboration, and communication (Nielsen, 2014). However, there is limited research on the features of effective internal feedback to provide empirical evidence regarding relationships between writing processes and outcomes (Graham, 2018; Nielsen, 2014). Van der Kleij (2024) also highlighted the need for more studies on the nature of internal feedback generated from different sources and reference frames and the processing of this feedback in specific contexts. This study evaluates internal feedback in terms of its amount, content, and accuracy, which have been considered feedback quality indicators in prior research on peer feedback (Zhang & Schunn, 2023; Zhang et al., 2024; Zong et al., 2021).

The amount of internal feedback can indicate the number of idea units the students generate in their feedback or the length of comments (Lu et al., 2021; Zhang et al., 2024). The students providing a higher amount of feedback for peers were found to have higher perceptions of their feedback helpfulness (Zong et al., 2021), be more likely to revise their own writing, and gain greater development in their writing ability (Cho & MacArthur, 2011; Wu & Schunn, 2021). However, from the perspective of the feedback receiver, more feedback comments have been linked to both increased (Wu & Schunn, 2021) and decreased feedback implementation (Patchan et al., 2016). Regarding internal feedback, Lu et al. (2021) also found a negative impact of feedback amount on students' writing improvement.

The content of internal feedback can be another feedback quality indicator with two core components: evaluative and elaborated (Narciss, 2008). While the evaluative component reveals the performance level, the elaborated component provides additional information about the topic, task, errors, and solutions. In the following classifications of the feedback content, four main categories have been frequently addressed, especially in writing tasks: summary, praise, problem, and solution (Lu et al., 2021; Nelson & Schunn, 2009; Patchan et al., 2016). Each of them can have a different role in and impact on feedback processing and improving writing (Lu et al., 2021). For example, Patchan et al. (2016) found that praise increased the likelihood of implementing the received peer feedback but did not improve writing quality. In the study by Nelson and Schunn (2009), summary and solution comments were positively correlated with feedback receivers' problem understanding in their writings; however, only providing solutions was directly

related to implementing feedback. Cho and MacArthur (2011) also indicated positive correlations between students' problem detection and suggestions for peers' works and their writing quality. In regard to internal feedback, there is limited research addressing the type of feedback content students provide to themselves (Panadero et al., 2019).

The accuracy of internal feedback as a feedback quality indicator can be categorized into scoring accuracy for the scores students give themselves and content accuracy for the elaborated feedback components (León et al., 2023; Panadero et al., 2016). Scoring accuracy in this study refers to the extent of alignment between students' and experts' (e.g., teachers and researchers) scorings by acknowledging the possibility of unreliable expert ratings (Andrade, 2019). Several reviews and meta-analyses on self-assessment (e.g., Andrade, 2019; Brown et al., 2013; Falchikov & Boud, 1989; León et al., 2023) revealed that the consistency between self-scorings and external judgments (e.g., teacher grades) was not high despite a considerable overlap, and students overestimated their performance. Most studies in the self-assessment field analyzed scoring accuracy; on the other hand, content accuracy, which is another but probably more critical aspect of feedback accuracy, has received limited research attention (Andrade, 2019; Peters et al., 2018; Panadero et al., 2016, 2019). Content accuracy refers to the extent of alignment between students' and experts' feedback comments on correct and incorrect aspects and improvement strategies for student performance (Panadero et al., 2016). Although both types of feedback accuracy are essential for learning, describing strengths or weaknesses accurately in self-assessment can be more beneficial for understanding ways of improving and regulating the learning process and performance than scoring accuracy (León et al., 2023).

It is also essential to note that generating content-rich and accurate feedback might not be achieved by all students, especially those with low knowledge and expertise in the subject area, and in complex tasks (e.g., writing), as indicated in previous research (e.g., Falchikov & Boud, 1989; León et al., 2023; Patchan & Schunn, 2015; Radović et al., 2024) and theoretical frameworks (e.g., Narciss, 2017; Butler & Winne, 1995; Nicol, 2021). To mitigate the effect of individual and contextual conditions on internal feedback generation, there is a need for strong support structures (Panadero et al., 2016). Scaffolding self-assessment might be more critical, especially in distance education and massive open online courses where students are less likely to get immediate teacher feedback. Accordingly, this study investigates the impact of an under-researched support strategy in an online learning environment: peer feedback generation on quality and poor writing examples. It can help students improve their mental structures about the task and, in turn, affect students' choices regarding the content and organization of their subsequent writings (Andrade, 2019).

Effects of exemplars on self-assessment and performance

Exemplars and peer works are two types of external information learners can use to compare their performance with them and generate internal feedback (Nicol, 2021). Information obtained from these sources can help students calibrate their self-assessments (Narciss, 2017; Yan & Brown, 2017), function as proxies for teacher feedback by increasing student agency in the feedback process, and mitigate teacher workload (Carless, 2022).

Concerning exemplars, it is recommended to use exemplars of different qualities and encourage learners to compare and contrast them in order to enhance their understanding of assessment criteria, feedback generation, and performance improvement (Narciss, 2017). However, empirical studies provide conflicting results about the impact of the quality of exemplars. For example, in the group discussions on utilizing exemplars for revising research reports, college students in Lipnevich et al.'s (2014) study explained that they primarily benefitted from the excellent exemplar, not weak or average exemplar because they regarded excellent exemplar as an effective, efficient, and easy way to improve their writing. On the other hand, some empirical studies favor the combined use of high-quality and low-quality examples rather than only high-quality ones. For instance, Lin-Siegler et al. (2015) investigated the effects of teachers' discussion of either contrasting examples (well and poorly-written stories) or only well-written examples. They found that the group with contrasting cases had higher quality stories, a deeper understanding of assessment criteria, more feedback comments about their weaknesses and improvement strategies, and scoring accuracy than the group with only well-written examples. Jax et al. (2019) also produced similar results in the study on high school students' performance and self-assessment accuracy in the physics course by favoring the group that compared their works with contrasting examples rather than only good examples.

Effects of peer feedback generation on self-assessment and performance

Peer works on which students generate peer feedback can also serve as examples of successful and failed performance (Peters et al., 2018). Therefore, assessing and generating feedback for peer works might contribute to internal feedback generation and performance improvement as well. Correspondingly, Cho and MacArthur (2011) found that the group reviewing peers' laboratory papers in the physics course outperformed the group only reading (un)related materials in the writing quality of these papers. Similarly, Wu et al. (2022) revealed that university students reviewing peers' business letters with varying qualities had significantly higher revised performance and transfer scores than those reviewing only their own letters. Qualitative studies (Nicol et al., 2014; To & Panadero, 2019) also corroborated the benefits of generating peer feedback for students to comprehend the assessment standards better, reflect on their work, more objectively judge their work, and revise their work even before they received feedback from their peers.

However, some studies failed to find the positive impact of peer feedback generation on performance improvement or internal feedback. For instance, Peters et al. (2018) revealed no significant differences in terms of the quality of vocational students' revised performance in a planning task and their internal feedback on missing and faulty parts in their performance and suggestions for improvement. In another study by Nicol and McCallum (2022), undergraduate students also generated internal feedback after reviewing each peer's essay by comparing their and peers' essays. Concerning self-assessment accuracy, the researchers revealed that almost all students, independent of their ability levels, determined the same areas for improvement in the essays as their teacher. However, it was also found that most students could accurately rank the essays they reviewed in terms of quality but not their own.

Considering the contradictory findings about the learning benefits of peer feedback generation, it is necessary to indicate the impact of contextual and learner factors that might affect these results (van Popta et al., 2017). The researchers (Cho & Cho, 2011; Nicol & McCallum, 2022) recommended taking into account the effect of the quality of assessed peer works. However, the existing evidence in this regard has also remained inconclusive in suggesting a specific quality level. To illustrate, Nicol and McCallum (2022) found that students at a medium level of essay writing competency benefitted from reviewing both lower- and higher-quality peer essays with descriptive analysis, although their internal feedback content changed based on the quality of peer essays. In Wu et al.'s (2022) study, students' comments on the weaknesses of their peer writings at the medium quality level significantly predicted both their revised writing performance and total transfer scores. However, Dmoshinskaia et al. (2022) favored secondary education students' reviewing of concept maps at low quality to achieve higher post-test scores, whereas Alqassab et al. (2018) indicated that providing feedback on the near-correct peer solution for a geometry proof led to higher proof comprehension than providing feedback on erroneous peer solution. In contrast to these studies, Cho and Cho (2011) did not find a direct effect of the peer-work quality on students' writing revision for the laboratory reports in a university physics course.

Overall, there is a need for further experimental studies that can elucidate which quality of peer works in peer assessment can be effective for follow-up self-assessment and performance improvement. Such an investigation can contribute to the literature that warrants more empirical evidence about whether and when the combined use of two assessment methods is more effective (Cheong et al., 2023).

Research aim and questions

This study aimed to investigate the effects of the quality of the assessed peer works (i.e., high, low, or both high and low qualities) or peer work quality (this shorter term is used henceforth) on writing performance and internal feedback. It also explored the possible costs of the intervention regarding time and mental effort allocated to peer feedback generation and examined the internal feedback features associated with writing performance. The following research questions guided this study:

(1) What are the effects of peer work quality on students' writing performance?
(2) What are the effects of peer work quality on the following features of internal feedback?

   (a) Amount
   (b) Content (i.e., summary, praise, problem, specific and general suggestion)
   (c) Problem detection accuracy (i.e., hit and false alarm rates)
   (d) Scoring accuracy (i.e., inaccuracy and bias scores)

Explorative analyses

(3) What are the effects of peer work quality on the peer feedback generation process (i.e., time on task and perceived mental effort)?

(4) What internal feedback features contribute to the prediction of the revised writing performance in addition to the initial writing performance?

Method

Participants and design

The inclusion criteria for the participants of this study consisted of (1) being a bachelor psychology student or a master's or doctoral student from all fields, (2) having English reading and writing skills at least B2 level according to the Common European Framework of Reference for Languages, (3) completing all tasks of the experiment, (4) providing writing and feedback generation tasks with original responses that were not generated by artificial intelligence (AI) or in short duration and did not exceed 20% plagiarism score, and (5) not getting the full point from the researchers' scoring of initial abstract.

A total of 211 students who met the first three inclusion criteria participated in this online experiment. However, after the originality of their responses was checked with the Copyleaks tool (https://copyleaks.com/), data from 47 participants (22.27%) were removed because they did not provide their own responses necessary for self-assessment. AI might have generated 100% of 25 participants' texts, considering their writing duration and word counts as well,[1] and the plagiarism criterion was violated by 22 participants. The final sample included 164 participants (126 female; $M_{age}=25.82$, $SD=4.71$). They were randomly assigned to one of the three experimental conditions differing in peer work qualities. One group gave peer feedback on two low-quality writings (PF−− , $n=55$), one group on one high-quality and one low-quality writing (PF+− , $n=57$), and one group on two high-quality writings (PF++ , $n=52$).

Materials

Online writing and assessment environment

The researchers developed an online writing and assessment environment for this study. Before implementing it in the study, its usability, design, and instructions were iteratively improved based on the suggestions of two user-experience experts and pilot studies with three undergraduate and three graduate students. It provided five tasks to learn and practice scientific abstract writing in English. These tasks were: (1) studying standards of a scientific abstract in terms of five assessment criteria, (2) writing an initial abstract for a study and self-scoring the abstract on a ten-point rating scale, (3) generating qualitative peer feedback on two fictitious students' abstracts labeled with high or low quality, (4) generating qualitative internal feedback and self-scoring, and (5) revising the initial abstract based on internal feedback and self-scoring.

[1] Of these participants, 17 might have used AI to generate an initial abstract. The average duration for reading the study and writing an abstract was 6.58 min ($SD=5.06$) for these participants, whereas it was 40.69 ($SD=29.57$) for those who generated themselves. There were also 11 participants who might have used AI to generate internal feedback. Although their writing duration for generating internal feedback ($M=4.80$, $SD=4.12$) was close to that of the participants who generated themselves ($M=5.71$, $SD=5.17$), the word count in the internal feedback of AI-group ($M=296.18$, $SD=184.88$) was substantially higher than that of the others ($M=77.61$, $SD=48.07$). After acknowledging the shortcomings of the accuracy of AI detector tools and considering participants' writing durations and word counts as well, the authors decided to remove data from 47 participants.

Abstract writing task

In the writing task, participants were first requested to read an article about the effect of cues of being watched on cooperation (Bateson et al., 2006) without an abstract. After reading, they were asked to write an abstract of no more than 250 words for this study. They were also warned not to use any external tool for writing and not to copy sentences from the study.

Fictitious peer abstracts

In the peer feedback stage, participants were first prompted to read and compare two fictitious peer abstracts presented side by side and indicated as high or low quality. Their qualities differed based on the intervention condition assigned to participants (i.e., PF−− , PF+− , PF++). The researchers prepared two low-quality and two high-quality abstracts for this stage. Typical errors in abstract writing were integrated into low-quality abstracts after reviewing the literature and interviewing with a consultant in the academic writing center at the authors' university. Low-quality abstracts were scored three out of ten by the independent ratings of the authors and a writing consultant, who utilized the rubric of this study to assess participants' writings. In contrast, high-quality ones got full points from all scorers.

Peer and internal feedback forms

The peer feedback form prompted students to fill out two input fields for each assessment criterion and think about similarities and differences between the assigned two abstracts (Fig. 1). The first input field was for describing and justifying the correct and incorrect aspects of each abstract. The participants were required to respond to them to proceed to the next page. The second field was for providing suggestions for improvement.

The internal feedback form also included the same input fields for participants to evaluate their own abstracts. In addition, it asked them to score their abstracts over 10, considering five assessment criteria. A self-scoring form was available after participants

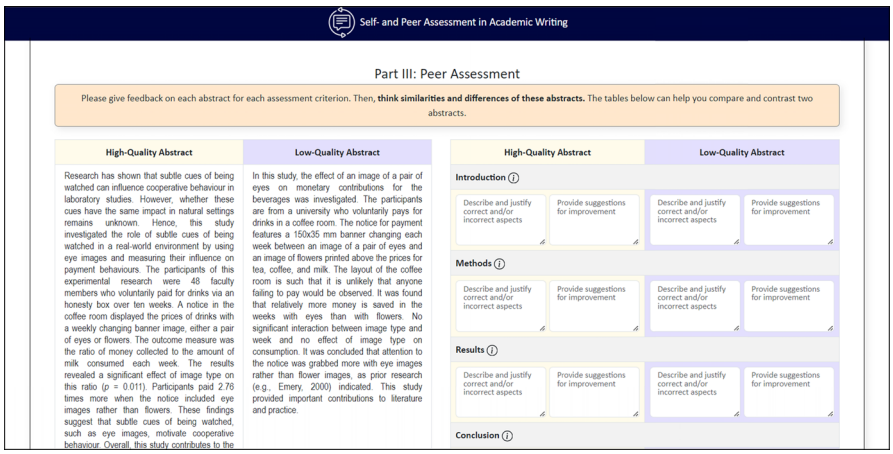


Fig. 1 Peer feedback form

wrote their initial abstract, generated qualitative internal feedback, and revised their writings.

Measures

Abstract writing performance

The abstract writing performance was assessed with a rubric (U.S. DOE Office of Science, 2011) adapted in this study. In addition to language and conventions as one criterion, it includes four criteria in line with the IMRD (Introduction–Methods–Results–Discussion) structure, the commonly used abstract writing structure in social sciences (Lorés, 2004). These criteria were also utilized in previous peer and self- assessment research (e.g., Krebs et al., 2022; Lu et al., 2021), but their rating scales differed. Three levels are described for five assessment criteria in the rubric of this study: strong (2 points), medium (1 point), and weak (0 points). Therefore, the total maximum score for an abstract can be 10. Original level descriptions were modified after the authors rated the sample abstracts with the rubric and discussed their appropriateness for the study. They ensured that the revised descriptions are explicit enough for the person who assesses an experimental research abstract in social sciences (See Supplementary Table S1).

The first author used this rubric to score both initial and revised abstracts. A graduate student assistant trained in using the rubric for 2 h also independently scored 25 initial and 25 revised abstracts randomly selected. Intraclass correlation coefficients were 0.86 and 0.89 for the initial and revised abstracts.

Internal feedback amount

The internal feedback of each participant was divided into comments based on idea units to determine the feedback amount. An idea unit refers to a comment that addresses one particular issue or topic in writing (Patchan et al., 2016). The length of the unit might change from a few words to a couple of sentences. A total of 1274 idea units were obtained after segmenting feedback into comments in this study.

Internal feedback content

The content of each idea unit was analyzed based on the coding scheme adapted from prior studies (Lu et al., 2021; Nelson & Schunn, 2009; Patchan et al., 2016). The scheme included six categories: summary, praise, problem, general suggestion, specific suggestion, and neutral. These categories closely match the coding scheme by Lu et al. (2021). They reported inter-rater reliability estimates equal to or higher than 0.70 for the analyses of peer feedback and self-feedback on abstract writing quality. Supplementary Table S2 describes these categories with sample feedback comments. Furthermore, Fig. 2 illustrates the feedback amount and content of idea units for one participant.

For this study, the inter-rater reliability coefficients were also calculated between the first author and the graduate student assistant who received training on coding feedback content and accuracy and analyzed each feedback comment of 25 participants independently. Kappa statistics for feedback content categories ranged from 0.51 (91.6%) to 0.89. Supplementary Table S2 provides Kappa measures for all feedback categories. The disagreements between the coders were also discussed and resolved.

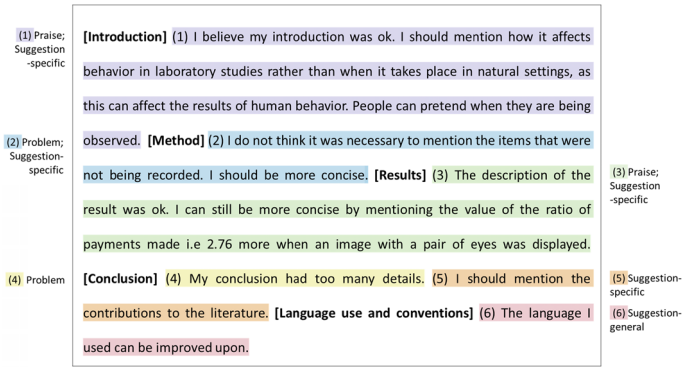


Fig. 2 Sample content analysis of feedback comments. Each color represents an idea unit that can include different feedback content types

Problem detection accuracy of internal feedback

To check the accuracy of problem detection in internal feedback, Wu and Schunn's (2023) procedure to analyze problem identification accuracy in peer feedback was used (Fig. 3). Wu and Schunn (2023) analyzed secondary school students' assessment performance in correctly determining problems in peers' essays in regard to seven common feedback topics after they selected implementable peer comments (e.g., problem statements) and coded their topics. Similarly, in our study, feedback comments, including problems and general or specific suggestions, were first chosen.

Second, these comments were coded to determine the topic of the problem addressed. A total of 27 common topics related to five assessment criteria were identified. Supplementary Table S3 lists these topics with sample feedback comments. The kappa statistics between the two coders for the problems addressed in internal feedback by 25 participants was 0.86.

Third, participants' initial abstracts were checked to determine their writing performance on the specified feedback topics at two levels: competent and problematic. While the competent level shows a well-written text, the problematic level suggests the need to improve the text to some/great extent. Then, the total number of problematic topics in the initial abstract of each participant was calculated.

Fourth, the problems addressed in feedback comments were compared with researchers' coding of problematic topics. Then, problem detection accuracy was coded based on signal detection theory (SDT) (Green & Swets, 1966) (See Supplementary Table S4). Hit and false alarm rates were calculated in Wu and Schunn's (2023) study on problem identification accuracy in peer feedback. The same measures were obtained in this study. Accurate comments correctly addressed a problem that existed in the initial abstract according to feedback topics, which is called "hit" based on SDT. On the other hand, inaccurate ones addressed an issue that was not actually problematic, which is called a "false alarm." There were also vague comments that stated a problem quite abstractly, so they were disregarded in further analyses. The intercoder reliability analysis was also conducted to categorize problems in regard to accuracy, which resulted in almost perfect agreement ($\kappa = 0.84$).

Finally, hit rates and false alarm rates were calculated according to SDT. The hit rate corresponds to the number of problems accurately detected by the participant (hit)

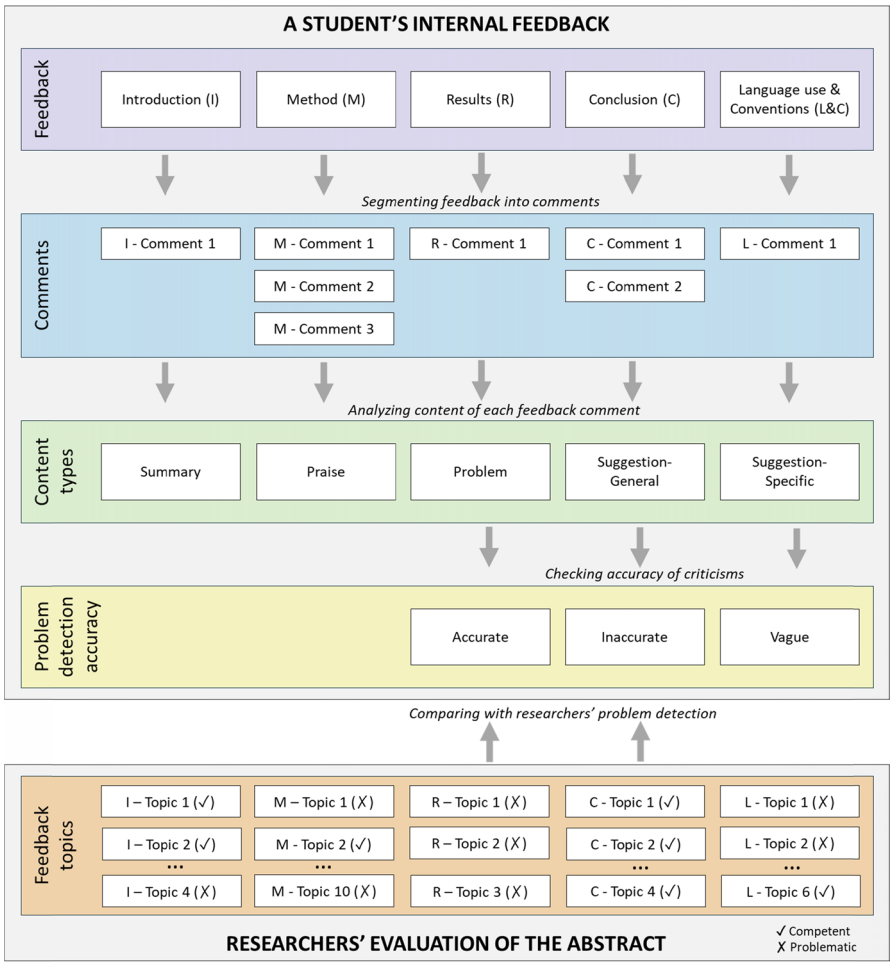


Fig. 3 The procedure for analyzing internal feedback

divided by the total number of problematic topics. The false alarm rate was obtained by dividing the false alarm number by the total number of unproblematic or competent topics in the participant's abstract. The whole feedback coding process is summarized in Fig. 3 by following a similar procedure in Patchan and Schunn's (2016) study.

Scoring accuracy

Based on previous research, two measures were calculated to assess scoring accuracy (e.g., Hosein & Harle, 2018; Krebs et al., 2022; Nietfeld & Schraw, 2002): inaccuracy and bias scores. Inaccuracy scores referred to the absolute difference between the students' self-scores and researchers' scores, but bias scores indicated the signed difference between the scorings of two agents ($X_{\text{student's self-score}} - X_{\text{researcher's score}}$). While positive bias scores imply students' overestimation of their performance, negative scores imply underestimation.

Time and perceived effort in peer feedback generation

The log data of the online environment was checked to obtain the total time in minutes each participant spent for peer feedback generation. Moreover, one question was used

to ask how much mental effort participants invested during peer feedback generation. It had a nine-point scale ranging from "(1): very, very low mental effort" to "(9): very, very high mental effort" (Paas, 1992). For the reliability of the scale, Cronbach's coefficient alpha was found to be 0.90 in Paas' (1992) study.

Entry characteristics

To check whether the intervention groups had similar background characteristics at the beginning of the study, participants responded to questions related to their current study level, program name, perceived English reading and writing skills according to the six-level descriptors of the Common European Framework of Reference for Languages, and writing experience (i.e., the total number of English writing assignments completed at the university). In addition, participants rated their confidence in (1) understanding an English research article published in a professional psychology journal and (2) writing an English abstract for this article from 1 to 100.

Procedure

After approval of the study from the university's ethical board, the volunteer university students completed the tasks of this online experiment with directions written in English in a self-paced and step-by-step manner (Fig. 4). They were discouraged to return to the previous steps. The average study completion time was about 1 h and 20 min. At the end of the study, study materials about standards of abstract writing and both high- and low-quality abstracts were shared with all participants.

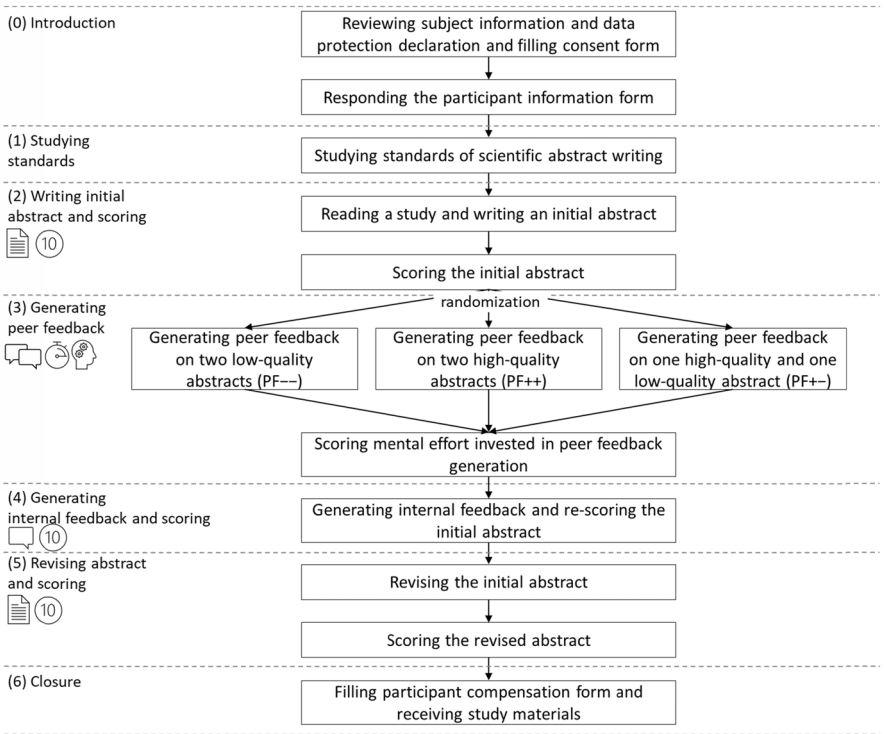


Fig. 4 The study procedure

Statistical analyses

Different statistical analyses were conducted in R version 4.3.3 to answer research questions. First, intervention groups' entry characteristics were compared to ensure that the groups were equivalent. The results (Supplementary Table S5) indicated that the groups did not differ significantly in regard to several variables explained in the *Entry characteristics* section.

To answer the research question regarding writing performance, mixed ANOVA was performed because of both between-subject (i.e., peer work quality: PF−− , PF+− , and PF++) and within-subject factors (i.e., time: initial and revised writing) as independent variables. Before that, its assumptions were checked. First, two extreme cases that received full points for their initial abstracts according to researchers' scoring were detected and removed since they were not considered members of the population who needed an intervention to detect problems in their abstracts and improve them. Second, the Shapiro–Wilk test indicated violations of normality assumption in these analyses. However, considering sample sizes greater than 50 in groups, the robustness of the F test to the violation of normality was expected (Tabachnick & Fidell, 2013). Third, Levene's tests and Box's M tests indicated that the assumptions of homogeneity of variance ($p > 0.05$) and variance–covariance matrices ($p > 0.001$) were met (Verma, 2016). It is also important to note that post-hoc comparisons for the significant main and interaction effects were evaluated with the Bonferroni-adjusted p values.

The effect of the intervention on the feedback amount was answered through ANOVA. On the other hand, MANOVAs were planned for the feedback content, problem detection accuracy, and scoring accuracy, which consisted of two or more related subcategories. The assumption checks for these analyses revealed one multivariate outlier, an unusual combination of variables concerning the feedback content based on Mahalanobis distance scores. Since the deletion of this outlier did not change the main results, it was retained. Other assumptions of MANOVA were met for the analyses regarding feedback content and problem detection accuracy.

However, the assumption of homogeneity of variance and covariances was violated in the analyses regarding scoring accuracy, specifically inaccuracy score. Instead of conducting a mixed MANOVA, separate mixed ANOVAs were performed for bias and inaccuracy scores, with peer work quality as a between-subject variable and time (i.e., after initial writing, feedback generation, and revised writing) as a within-subject variable. Robust mixed ANOVA was performed to examine the main and interaction effects for the inaccuracy scores. In addition, the sphericity assumption was violated in the analysis of bias scores. Since Greenhouse–Geisser epsilon was higher than 0.75, the Huynh–Feldt correction was used (Girden, 1992).

For the explorative analyses concerning the effects of the intervention on the peer feedback generation process, the robust ANOVAs based on trimmed means (20% trimming) (Mair & Wilcox, 2020) were conducted because of one extreme outlier for mental effort and five outliers for peer feedback generation duration. Finally, stepwise regression analysis was applied as an explanatory technique (Tabachnick & Fidell, 2013) to build a model with feedback features that significantly contributed to the prediction of the revised abstract quality. The model did not include two feedback features (i.e., amount and inaccuracy) because of the multicollinearity problem emerging

from their high correlations with other predictors. The two models that included only initial writing quality (model 1) and additionally significant internal feedback features (model 2) were also compared to check the improvement in the model with these feedback features.

Findings

The effects of peer work quality on students' writing performance

Descriptive statistics concerning all variables are listed in Table 1. The mixed ANOVA yielded a significant main effect of time and peer work quality and their significant interaction effect on writing performance (Table 2). Post-hoc comparisons revealed that the revised writing performance was significantly higher than the initial writing performance for all three conditions: PF−− ($p = 0.006$), PF+− ($p < 0.001$), and PF++ ($p < 0.001$). In addition, there was a nonsignificant difference between the conditions in the initial writing performance ($p = 0.842$), but a significant difference occurred in the revised writing performance ($p = 0.002$). The PF++ ($p = 0.001$) and PF+− ($p = 0.014$) groups had significantly higher revised writing performance than the PF−− group.

Table 1 Descriptive statistics regarding writing performance, internal feedback, and peer feedback

Measure	PF−− ($n = 54$)		PF+− ($n = 56$)		PF++ ($n = 52$)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Writing performance						
Initial performance	4.17	1.46	4.36	1.41	4.54	1.49
Revised performance	4.52	1.61	5.46	1.67	5.75	1.90
Feedback amount						
Idea unit	7.20	2.73	8.30	2.78	7.85	2.61
Feedback content						
Summary	2.00	2.54	2.16	2.10	1.88	2.31
Praise	2.17	1.61	2.12	1.45	2.25	1.71
Problem	1.80	1.63	2.75	1.71	2.94	1.90
General suggestion	1.80	1.35	1.48	1.36	1.17	1.31
Specific suggestion	1.67	1.59	2.96	2.07	2.67	2.02
Problem detection accuracy						
Hit rate	0.19	0.16	0.32	0.16	0.29	0.18
False alarm rate	0.05	0.06	0.07	0.06	0.07	0.06
Scoring accuracy						
Inaccuracy after the initial writing	2.52	1.63	2.18	1.47	1.75	1.25
Inaccuracy after feedback generation	2.32	1.68	1.86	1.18	1.38	1.01
Inaccuracy after revised writing	3.04	1.95	2.16	1.51	1.65	1.24
Bias after the initial writing	2.26	1.98	1.93	1.79	1.40	1.64
Bias after feedback generation	2.02	2.03	1.39	1.71	0.19	1.72
Bias after revised writing	2.85	2.22	1.80	1.93	0.85	1.89
Peer feedback generation						
Duration	16.52	14.70	20.41	20.44	15.8	9.58
Mental effort	6.65	1.23	6.50	1.03	6.31	1.23

Table 2 Inferential statistics regarding writing performance, internal feedback, and peer feedback

Variable	Pillai's Trace	*df*	*F*	*p*	η_p^2
Writing performance					
Time		1159	106.02	< 0.001	0.400
Peer work quality		2159	4.04	0.019	0.048
Interaction		2159	9.78	< 0.001	0.110
Feedback amount		2159	2.27	0.106	
Feedback content	0.17	10,312	2.88	0.002	0.08
Summary		2159	0.19	0.824	
Praise		2159	0.09	0.918	
Problem		2159	6.61	0.002	0.077
General suggestion		2159	2.86	0.06	
Specific suggestion		2159	6.96	0.001	0.08
Problem detection accuracy	0.10	4318	2.88	0.003	0.05
Hit rate		2159	8.47	< 0.001	0.096
False alarm rate		2159	1.27	0.283	
Inaccuracy scores					
Time		282.47	7.04	0.002	0.146
Peer work quality		258.95	7.62	0.001	0.205
Interaction		464.90	1.05	0.386	
Bias scores					
Time		1.88298.53	21.47	< 0.001	0.119
Peer work quality		2159	11.37	< 0.001	0.125
Interaction		3.76298.53	5.29	0.001	0.062
Peer feedback generation					
Duration		26375	0.131	0.877	
Mental effort		262.82	0.796	0.456	

df degrees of freedom; *F* *F*-ratio, test statistics from (mixed) (M)ANOVA; *p* probability value; η_p^2 partial eta squared as an estimate of effect size

The effects of peer work quality on internal feedback

Amount

The internal feedback amount was comparable among conditions, indicating that peer work quality did not affect the amount (Table 2).

Content

The MANOVA results showed a significant multivariate effect of the peer work quality on the feedback content. After adjusting the *p*-value with Bonferroni correction (0.05/5 = 0.01), follow-up univariate ANOVAs indicated that peer work quality had no significant effect on the number of summaries, praises, and general suggestions in internal feedback (Table 2). However, it significantly affected the number of problems and specific suggestions. Post-hoc Tukey tests revealed that the PF++ ($p = 0.004$) and PF+− ($p = 0.009$) groups stated significantly more problems in internal feedback than the PF−− group, but there was no significant difference between PF++ and PF+− groups ($p = 0.846$). Moreover, there were higher numbers of specific suggestions in the PF++ ($p = 0.015$) and PF+− ($p = 0.001$) groups compared to the PF−− group; however, the difference between PF++ and PF+− groups was nonsignificant ($p = 0.74$).

Problem detection accuracy

The MANOVA results also revealed a significant multivariate effect of peer work quality on problem detection accuracy (Table 2). Univariate ANOVAs indicated a significant effect on the hit rate but not on the false alarm rate. The hit rate of the PF+− ($p<0.001$) and PF++ ($p=0.011$) groups was significantly higher than the PF−− group, although the PF+− and PF++ groups did not differ significantly ($p=0.717$).

Scoring accuracy

The results of robust mixed ANOVA for inaccuracy scores showed a significant effect of time and peer work quality but a nonsignificant interaction effect (Table 2). In regard to the time factor, the participants were more accurate at scoring their initial writings after feedback generation ($p=0.021$), but their accuracy decreased after that when they revised their writings ($p<0.001$). There was no significant difference between inaccuracies after initial and revising writings ($p=0.741$). Concerning the peer work quality, the PF++ group gave the most accurate scores themselves compared to the PF−− group ($p<0.001$) and PF+− ($p=0.013$). The PF+− was also more accurate than the PF−− group ($p=0.002$).

The results for bias scores also yielded a significant main effect of time and peer work quality (Table 2). There was also a significant interaction effect. Across time, post-hoc comparisons revealed that bias scores significantly changed for all three groups: PF−− ($p<0.001$), PF++ ($p<0.001$), and PF+− ($p=0.033$). According to pair-wise comparisons, participants' bias for their initial writing performance significantly decreased in the groups of PF++ ($p<0.001$) and PF+− ($p=0.019$) after feedback generation. However, then, bias scores rose in all groups (PF−− : $p<0.001$, PF++ , $p=0.008$; PF+− , $p=0.016$) after they revised their writings. Moreover, there was a significant increase in bias after the initial and revised writing in the group PF−− ($p=0.01$).

Across experimental groups, bias scores did not significantly differ after initial writing ($p=0.156$) but differed after generating feedback ($p<0.001$) and revising writings ($p<0.001$). The bias score of the PF++ group was significantly less than that of PF−− ($p<0.001$) and PF+− ($p=0.002$) groups after feedback generation (Fig. 5). Similarly, the PF++ group had less bias than the PF−− ($p<0.001$) and PF+− ($p=0.045$) groups after

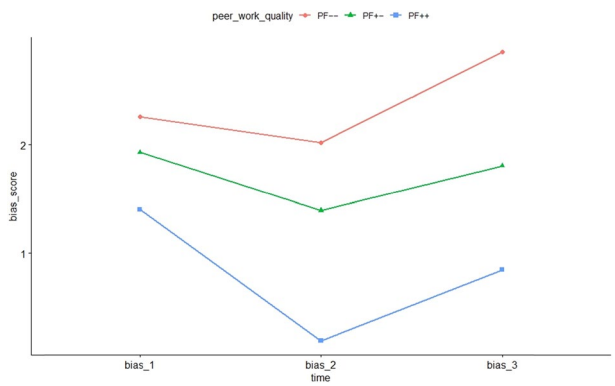


Fig. 5 Marginal means plot of time X peer work quality for bias scores. PF−− = Peer feedback on two low-quality peer writings, PF+− = Peer feedback on one high-quality and one low-quality writing, PF++ = Peer feedback on two high-quality writings

revised writing. The difference between the PF−− and PF+− groups was also significant after revised writings ($p = 0.022$), favoring the latter for less bias.

Exploratory analyses

The first explorative analysis sought to determine the impacts of peer work quality on the peer feedback generation process. Robust one-way ANOVA results revealed that peer work quality did not significantly affect peer feedback generation duration and perceived mental effort (Table 2).

Concerning correlations between writing performance and internal feedback features, revised abstract quality was significantly and largely correlated with hit rate and bias score after feedback generation and moderately with feedback amount, specific suggestions in internal feedback, and inaccuracy (Table 3). Regression analyses were also conducted to explore the internal feedback features that contributed to predicting the revised writing performance by taking into account the initial writing performance. The first model included only the initial writing performance and significantly predicted the revised writing performance (Table 4). Internal feedback features were also added to the second model, and stepwise regression with stepwise selection was performed. Hit rate and bias score after peer feedback generation significantly contributed to the model in addition to the initial writing performance (Table 4). Those who correctly detected problems and had less bias in self-scoring after feedback generation were more likely to have higher revised writing performance. In addition, the comparison of the fit of the first and second models revealed that adding the two internal feedback features led to a significant improvement in the model, $F\ (2,\ 158) = 18.77$, $p < 0.001$. More specifically, the hit rate and bias score in the second model accounted for the additional 7.95% variance in the revised writing performance beyond the variance explained by the initial writing performance.

Discussion

The effects of peer work quality on students' writing performance

This study revealed the significant effect of peer work quality on writing performance. Specifically, all groups improved their writing performance after feedback generation; however, the students assessing two high-quality peer works or one high- and one low-quality work had higher revised writing performance than those assessing two low-quality peer works. This study first provides empirical evidence that complementing peer assessment with formative self-assessment has beneficial effects on university students' writing performance.

Moreover, it adds new empirical insights into the interplay of formative self- and peer assessment by drawing attention to the quality of the assessed peer work. In a previous study (Cho & Cho, 2011), the direct effect of peer work quality on students' writing revision was not significant. Furthermore, Nicol and McCallum (2022) concluded that reviewing both low- and high-quality peer works could be beneficial for learning, especially for students with a medium-level writing competency, but they did not measure and compare to what extent participants improved their writing after reviewing each peer work with inferential analyses. Other relevant studies provided conflicting findings. They supported students' reviewing of low-quality (Dmoshinskaia et al., 2022),

Table 3 Correlations between writing performance and internal feedback features

	1	2	3	4	5	6	7	8	9	10	11
1. Initial writing performance											
2. Amount	0.26***										
3. Summary	0.16*	0.69***									
4. Praise	0.15	0.19*	− 0.05								
5. Problem	0.07	0.32***	− 0.01	− 0.20**							
6. General suggestion	− 0.23**	− 0.13	− 0.19*	− 0.11	0.06						
7. Specific suggestion	0.25***	0.56***	0.25**	− 0.03	0.21**	− 0.32***					
8. Hit rate	0.34***	0.60***	0.26***	− 0.06	0.41***	− 0.14	0.64***				
9. False alarm rate	0.27***	0.46***	0.08	0.10	0.28***	0.03	0.48***	0.22**			
10. Bias after feedback generation	− 0.46***	− 0.21**	− 0.03	0.21**	− 0.41***	0.12	− 0.26***	− 0.39***	− 0.17*		
11. Inaccuracy after feedback generation	− 0.29***	− 0.21**	− 0.10	0.19*	− 0.33***	0.11	− 0.19*	− 0.27***	− 0.10	0.69***	
12. Revised writing performance	0.77***	0.37***	0.23**	0.04	0.23**	− 0.22**	0.38***	0.51***	0.26***	− 0.50***	− 0.33***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 4 Regression results predicting the revised writing performance with initial writing performance and internal feedback features

	Model 1				Model 2			
Effect	***B***	***SE***	***β***	***t***	***B***	***SE***	***β***	***t***
Intercept	1.12	0.29	–	3.86***	1.32	0.34	–	3.94***
Initial abstract	0.95	0.06	0.77	15.05***	0.77	0.07	0.63	11.85***
Hit rate	–	–		–	2.55	0.53	0.25	4.82***
Bias after peer feedback generation	–	–		–	− 0.11	0.05	− 0.12	− 2.27*
Model fit	F (1, 160) = 226.5, R^2 = 0.59, adj. R^2 = 0.58, p < 0.001				F (3, 158) = 104.8, R^2 = 0.67, adj. R^2 = 0.66, p < 0.001			

*p < 0.05, **p < 0.01, ***p < 0.001

medium-quality (Wu et al., 2022), or high-quality (Alqassab et al., 2018) peer products to enhance their learning. The results of this online experiment conducted with university students who initially scored about four over ten in an abstract writing task favor the feedback generation on at least one high-quality peer work for the development of writing performance more substantially. Learning from only low-quality peer works or erroneous examples might be a more advantageous condition for students with high prior knowledge or skills, as indicated in previous studies (e.g., Große & Renkl, 2007; Heemsoth & Heinze, 2014). High-quality peer writings in this study might have been exemplars for the students to learn characteristics of well-written abstracts and apply them in their writings.

In addition, this study revealed no significant difference between the students comparing and generating peer feedback on two high-quality and one high- and one low-quality peer work, in contrast to some prior research results that supported the use of contrasting examples (e.g., Jax et al., 2019; Lin-Siegler et al., 2015). In these studies, the teacher facilitated the comparison of contrasting examples (Lin-Siegler et al., 2015), and participants compared the examples with their works (Jax et al., 2019). On the other hand, the current study presented contrasting examples simultaneously and prompted students to compare them and generate peer feedback. Even though the groups varying in the peer work quality did not differ in terms of perceived mental effort and duration in peer feedback generation, this type of presentation and comparison in complex tasks can be more effective for learners with high prior knowledge, as Meier et al. (2024) revealed.

The effects of peer work quality on internal feedback

The peer work quality also had a significant impact on the content and accuracy of internal feedback. Although it did not influence the amount of internal feedback and the number of summaries, praises, and general suggestions, the frequencies of problems and specific suggestions in internal feedback differed significantly by favoring the PF++ and PF+− groups. Internal feedback can change based on the reference information learners use for the comparison (Nicol, 2021). In this study, peer works were presented as a potential reference point for students, and it was revealed that their quality affected the content of internal feedback. Similarly, Nicol and McCallum (2022) found that the internal feedback of students assessing peers' essays at a lower quality than theirs was more likely to address the strengths in their essays, but it included more suggestions for

improvement when the assessed essays were of higher quality. Detecting problems and proposing solutions for major issues in writing require more subject and task knowledge than finding strengths in writing (Cho & Cho, 2011). Analyzing exemplars with different qualities is recommended to help students develop this knowledge (Narciss, 2017; Jax et al., 2019). Results of this study also specify that assigning at least one high-quality peer work to the students for peer assessment might contribute more to this knowledge and internal feedback generation by increasing the number of problem identifications and specific suggestions in internal feedback comments.

Concerning the accuracy of internal feedback, the findings also revealed that problem detection accuracy, particularly the hit rate, was significantly higher in the PF++ and PF+− groups than in the PF−− group. Generating accurate internal feedback is critical because inaccuracies in students' descriptions of their strengths and weaknesses in feedback cause them to follow the wrong ways to enhance their performance and regulate their learning (Kostons et al., 2009). However, students generally might not find all the problems in their writings accurately, and Wu and Schunn (2023) called for research on the factors that can influence assessors' problem detection accuracy. This study found that the quality of the assessed peer works affected the problem detection accuracy by favoring the groups reviewing at least one high-quality writing.

Finally, the scoring accuracy was compared between the three intervention groups of this study. The overall results concerning the bias and inaccuracy scores indicated that the PF++ group had the most accurate scores both after generating feedback and revising writings. The PF+− group was also more accurate than the PF−− group. The intriguing finding was that the bias of the PF−− group significantly increased from the initial to final writing. Accurate self-assessment as an essential attribute of effective metacognitive monitoring might not be easily attained (Winne & Azevedo, 2014), considering the common problem of students' overestimation of their performance (León et al., 2023). Therefore, there is a need for support structures that can help students compare their self-assessments with some reference information and achieve calibration between their judgments and actual performance (Narciss, 2017; Jax et al., 2019; Panadero et al., 2016). Social comparisons between students' and peers' performance might influence their calibration judgments in addition to internal, student-centered attributional sources (e.g., time and effort allocated for the task) (Bol & Hacker, 2012). However, social cues informing students of peers' task performance can have both positive and negative effects on the accuracy of judgments depending on the peer's performance level and students' metacognitive ability (e.g., de Carvalho Filho & Yuzawa, 2001). Results of this study also imply that having at least one high-quality peer work as a reference information or basis for calibration might enhance university students' knowledge about writing standards and scoring accuracy, whereas assessing only low-quality peer work might exacerbate the bias problem.

Exploratory analyses regarding the correlations between revised writing performance and internal feedback features

This study also explored relationships between revised writing performance and the amount, content, and accuracy of internal feedback. While the revised abstract performance was largely correlated with the hit rate in problem detection accuracy and bias

in self-scoring, medium correlations were found with feedback amount, specific suggestions in internal feedback, and inaccuracy in self-scoring. Further analyses also revealed that hit rate and bias score were the internal feedback features that best predicted the revised writing performance. Literature (Andrade, 2019; Panadero et al., 2016; Winne & Azevedo, 2014) highlights the need for high-quality research studies that can elucidate the relationship between students' formative self-assessment accuracy, successive study and revision behaviors, and learning. Concerning peer assessment, Cho and MacArthur (2011) found that assessors' writing quality was positively correlated with prior writing skills, the number of comments, problems, and suggestions in peer feedback. However, they revealed in regression analysis that problem detection was only a significant predictor of writing quality after peer reviewing. In He and Go's (2023) study, feedback accuracy or the accurate identification of problems was the significant positive predictor of students' implementation of peer feedback, with the largest effect size among all variables regarding feedback features and focus. Similarly, in the context of self-assessment, this study indicates the importance of feedback comments addressing problems in writing but also highlights accurate problem detection in internal feedback and less bias in self-scoring for learners to attain higher performance improvements.

Limitations and future directions

This online experiment has some limitations. First, it did not include a control group that did not generate peer feedback since we aimed to investigate the impact of the quality of the assessed peer works in this study. However, considering some empirical studies (e.g., Peters et al., 2018) that revealed the null effect of peer feedback generation on internal and performance improvement, future experimental studies can also involve a control group to discuss whether findings from qualitative studies about the benefits of peer assessment for self-assessment (e.g., Nicol et al., 2014; To & Panadero, 2019) hold true in randomized controlled trials.

Second, peer writings were labeled as high and low-quality during peer feedback generation in this study. However, peer assessment activities might be less likely to explicitly indicate the quality level of assessed works. Hence, prospective studies can examine whether the results of this study are transferable to peer assessment settings that do not show quality levels.

Third, the peer abstracts that the participants were explicitly prompted to compare were written for the same study. According to Nicol (2021), they can be called near comparators. Far comparators with different content but similar structures (e.g., abstracts written for a different study in this case) can be used in future studies to analyze the impact of different types of comparators on internal feedback generation and performance.

Fourth, the comparison prompts in this study were limited to the peer feedback generation stage. They might also be used in the internal feedback generation so that learners can compare their work with peer work. An analysis of internal feedback that stems from such a comparison can contribute to the scarce literature about the knowledge students construct during such a comparison (Nicol & McCallum, 2021).

Fifth, the percentages of accurately detecting existing problems in writing (i.e., hit rate) were not high; they ranged from 19 to 32%. This underscores the urgent need for studies

focusing on the accuracy of qualitative feedback and potential ways to increase problem detection accuracy, as this study emphasizes its significant contribution to the revised performance quality.

Sixth, although the authors collected data from 211 students initially, data from 47 participants (i.e., 22.27% of the collected data) were removed from this online experiment because of plagiarism and the possibility of using AI to generate responses. This problem caused the use of data from a smaller sample. Future studies might collect data from a larger sample to validate the findings of this study in another context. Moreover, the problem of collecting unoriginal responses from students warrants considerable attention from the researchers and educators who design online assessment tasks. Instead of prohibiting the use of AI tools in these tasks, future studies might design learning tasks that require learners to compare their own writings with AI-generated text and give feedback on these texts.

Seventh, the content of the generated peer feedback might also influence assessors' internal feedback generation and revision performance. Cho and Cho (2011) found that students' revised writings were positively affected by their comments addressing their peers' strengths for macro issues in writing and weaknesses for micro issues, although there was no direct effect of the peer work quality on students' own writing. In Wu et al.'s (2022) study, peer feedback comments on weaknesses seemed to better predict assessors' learning gains compared to comments on strengths. Future studies can provide further evidence about the indirect effect of the quality of the assessed peer works on internal feedback generation and writing via the features of the generated peer feedback.

Finally, other reference frames for learners to generate internal feedback can be used in future research. For example, in a recent study by Nicol and Kushwah (2024), three university students first compared two published literature reviews and then compared them with their writings. Their internal feedback generated from these comparisons closely matched instructor feedback. Quantitative studies with larger samples can provide further empirical evidence to support internal feedback generation through comparisons with external information. Digital environments can present different resources for this purpose (e.g., online resources, lecture presentations or videos, and AI-generated writings).

Practical implications

This study also has several practical implications. First, the results indicate improvements in university students' writing performance after online self-assessment complemented with peer assessment. As a result, the current research suggests the combined use of two assessment methods for formative purposes in the higher education context.

Another significant implication of the study is related to the allocation of peer works in peer assessment. Peer work is generally randomly assigned to the students by software without considering their quality (Nicol, 2021). Instead, the results of this study favor the review of at least one high-quality work to enhance internal feedback generation and writing performance. Accordingly, peer assessment systems can allocate at least one high-quality peer work to the reviewers or assessors. Automatic essay scoring supported by machine learning or AI can be used to identify high-quality writings. Among the recent discussions about the use of AI to enhance peer assessment and promote peer

feedback quality (e.g., Bauer et al., 2023; Topping et al., 2025), this study draws attention to the peer work quality as one factor that needs to be considered while assigning assessors so that they also benefit more from the feedback provision.

This study also reveals the internal feedback features significantly associated with revised writing performance. Results highlighted problem detection accuracy and bias as significant predictors of revised writing performance. Therefore, it is important for students to accurately identify their weaknesses without overestimating their current performance to improve their writing. Some support structures (e.g., prompts and training) can be designed to facilitate this process, and this study recommends learners' peer feedback generation on two high-quality or one high-quality and one low-quality peer works.

Finally, data from the participants who had higher than 20% plagiarism score or might have used AI to generate writing and feedback were removed in this study. This problem was encountered even though the participants received instructions about not using any external tools for writing and not copying sentences from the study they read during the online experiment. Therefore, it might be necessary to design more systematic interventions that inform higher education students about the good scientific practices and ethical use of AI (Kovari, 2025). This might be important, especially for the assessment tasks (e.g., abstract writing in this study) that AI can easily perform.

Conclusion

Generating quality internal feedback that involves a change process in conceptual, procedural, or metacognitive knowledge (Nicol & McCallum, 2022) is an important skill for becoming a self-regulated lifelong learner (Narciss, 2017) and persisting in online courses requiring student autonomy. However, the body of knowledge that elucidates the support mechanisms for internal feedback generation is limited (Andrade, 2019; Nicol et al., 2014), although technology-enhanced learning environments can facilitate the development and implementation of different self-assessment strategies (Radović et al., 2023). This study used the strategy of peer feedback generation and specified the impact of peer work quality on writing and internal feedback in an online learning environment. The results suggest that students who review at least one high-quality peer work can detect more problems in their own writing, propose more specific suggestions in their internal feedback, give more accurate self-scorings, and substantially improve their writing performance. In addition, the study highlights significant correlations between internal feedback features (i.e., problem detection accuracy and bias) and final writing performance. Overall, this study shows that combining formative peer and self-assessment is a promising strategy to promote writing performance, particularly if students receive at least one high-quality peer work.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41239-025-00522-4.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
EA contributed to the conceptualization, methodology, development of study materials, data collection and analysis, visualization, writing—original draft, review, and editing. SN contributed to the conceptualization, methodology, development of study materials, writing—review and editing, and supervision.

Funding
This research was funded by the Alexander von Humboldt Foundation within the scope of the postdoctoral fellowship granted to the first author. The second author is partly funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

Availability of data and materials
Some of the data analyzed during this study are included in this article and its supplementary file. The whole dataset is available from the corresponding author upon reasonable request.

Declarations

Competing interests
We have no conflicts of interest to disclose.

Received: 17 July 2024 Accepted: 11 March 2025
Published online: 11 April 2025

References

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). The impact of peer solution quality on peer-feedback provision on geometry proofs: Evidence from eye-movement analysis. *Learning and Instruction, 58*, 182–192. https://doi.org/10.1016/j.learninstruc.2018.07.003

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*, 87. https://doi.org/10.3389/feduc.2019.00087

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice, 48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*(3), 412–414. https://doi.org/10.1098/rsbl.2006.0509

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology, 54*(5), 1222–1245. https://doi.org/10.1111/bjet.13336

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology, 3*, 229. https://doi.org/10.3389/fpsyg.2012.00229

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom learning* (pp. 367–393). Sage Publications.

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

Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education, 23*(2), 143–153. https://doi.org/10.1177/1469787420945845

Cheong, C. M., Luo, N., Zhu, X., Lu, Q., & Wei, W. (2023). Self-assessment complements peer assessment for undergraduate students in an academic writing task. *Assessment & Evaluation in Higher Education, 48*(1), 135–148. https://doi.org/10.1080/02602938.2022.2069225

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73–84. https://doi.org/10.1037/a0021950

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643. https://doi.org/10.1007/s11251-010-9146-1

de Carvalho Filho, M. K., & Yuzawa, M. (2001). The effect of social influences and general metacognitive knowledge on metamemory judgments. *Contemporary Educational Psychology, 26*(4), 571–587. https://doi.org/10.1006/ceps.2000.1077

Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2022). Giving feedback on peers' concept maps as a learning experience: Does quality of reviewed concept maps matter? *Learning Environments Research, 25*(3), 823–840. https://doi.org/10.1007/s10984-021-09389-4

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

Girden, E. R. (1992). *ANOVA: Repeated measures*. Sage Publications.

Graham, S. (2018). Instructional feedback in writing. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 145–168). Cambridge University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction, 17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

He, W., & Gao, Y. (2023). Explicating peer feedback quality and its impact on feedback implementation in EFL writing. *Frontiers in Psychology, 14*, 1177094. https://doi.org/10.3389/fpsyg.2023.1177094

Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science, 42*, 639–657. https://doi.org/10.1007/s11251-013-9302-5
Hosein, A., & Harle, J. (2018). The relationship between students' prior mathematical attainment, knowledge and confidence on their self-assessment accuracy. *Studies in Educational Evaluation, 56*, 32–41. https://doi.org/10.1016/j.stueduc.2017.10.008
Jax, J., Ahn, J. N., & Lin-Siegler, X. (2019). Using contrasting cases to improve self-assessment in physics learning. *Educational Psychology, 39*(6), 815–838. https://doi.org/10.1080/01443410.2019.1577360
Kostons, D., van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology*, *23*(9), 1256–1265. https://doi.org/10.1002/acp.1528
Kovari, A. (2025). Ethical use of ChatGPT in education—Best practices to combat AI-induced plagiarism. *Frontiers in Education, 9*, 1465703. https://doi.org/10.3389/feduc.2024.1465703
Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning, 17*(2), 627–650. https://doi.org/10.1007/s11409-022-09302-1
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
Lin-Siegler, X., Shaenfield, D., & Elder, A. D. (2015). Contrasting case instruction can improve self-assessment of writing. *Educational Technology Research and Development, 63*(4), 517–537. https://doi.org/10.1007/s11423-015-9390-9
Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*(4), 539–559. https://doi.org/10.1007/s11251-013-9299-9
Lorés, R. (2004). On RA abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes, 23*(3), 280–302. https://doi.org/10.1016/j.esp.2003.06.001
Lu, Q., Zhu, X., & Cheong, C. M. (2021). Understanding the difference between self-feedback and peer feedback: A comparative study of their effects on undergraduate students' writing improvement. *Frontiers in Psychology, 12*, 739962. https://doi.org/10.3389/fpsyg.2021.739962
Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*, 464–488. https://doi.org/10.3758/s13428-019-01246-w
Meier, J. M., Hesse, P., Abele, S., Renkl, A., & Glogger-Frey, I. (2024). Video-based modeling examples and comparative self-explanation prompts for teaching a complex problem-solving strategy. *Journal of Computer Assisted Learning*, 1–19. https://doi.org/10.1111/jcal.12991.
Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. Van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.) (pp. 125–144). New York, NY: Lawrence Erlbaum Associates.
Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 173–189). Singapore: Springer Singapore.
Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*(4), 375–401. https://doi.org/10.1007/s11251-008-9053-x
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
Nicol, D., & Kushwah, L. (2024). Shifting feedback agency to students by having them write their own feedback comments. *Assessment & Evaluation in Higher Education, 49*(3), 419–439. https://doi.org/10.1080/02602938.2023.2265080
Nicol, D., & McCallum, S. (2022). Making internal feedback explicit: Exploiting the multiple comparisons that occur during peer review. *Assessment & Evaluation in Higher Education, 47*(3), 424–443. https://doi.org/10.1080/02602938.2021.1924620
Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518
Nielsen, K. (2014). Self-assessment methods in writing instruction: A conceptual framework, successful practices and essential strategies. *Journal of Research in Reading, 37*(1), 1–16. https://doi.org/10.1111/j.1467-9817.2012.01533.x
Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131–142. https://doi.org/10.1080/00220670209596583
Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429
Panadero, E., Brown, G. T., & 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, 31*, 221–253. https://doi.org/10.1080/0969594X.2024.2367027
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., Lipnevich, A., & Broadbent, J. (2019). Turning self-assessment into self-feedback. In D. Boud, M. D. Henderson, R. Ajjawi, & E. Molloy (Eds.), *The impact of feedback in higher education: Improving assessment outcomes for learners* (pp. 147–163). Palgrave Macmillan.
Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science, 43*, 591–614. https://doi.org/10.1007/s11251-015-9353-x
Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*(8), 1098–1120. https://doi.org/10.1037/edu0000103
Peters, O., Körndle, H., & Narciss, S. (2018). Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education*, *33*(1), 117–143. https://doi.org/10.1007/s10212-017-0344-y

Radović, S., Seidel, N., Haake, J. M., & Kasakowskij, R. (2024). Analysing students' self-assessment practice in a distance education environment: Student behaviour, accuracy, and task-related characteristics. *Journal of Computer Assisted Learning, 40*(2), 654–666. https://doi.org/10.1111/jcal.12907

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

To, J., & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assessment & Evaluation in Higher Education, 44*(6), 920–932. https://doi.org/10.1080/02602938.2018.1548559

Topping, K. J., Gehringer, E., Khosravi, H., Gudipati, S., Jadhav, K., & Susarla, S. (2025). Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education, 22*(1), 3. https://doi.org/10.1186/s41239-024-00501-1

U.S. Department of Energy, Office of Science. (2011). Research abstract evaluation rubric. https://dl.icdst.org/pdfs/files2/c20dc88a777d07244363b20bcf3edf2f.pdf. Accessed 10 July 2024.

Van der Kleij, F. M. (2024). Supporting students to generate self-feedback: Critical reflections on the special issue. *Studies in Educational Evaluation, 81*, 101348. https://doi.org/10.1016/j.stueduc.2024.101348

van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P.R.-J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review, 20*, 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

Verma, J. P. (2015). *Repeated measures design for empirical researchers* (1st ed.). Wiley.

Winne, P., & Azevedo, R. (2014). Metacognition and self-regulated learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 63–87). Cambridge University Press.

Wu, H., Zhu, H., & Yang, X. (2022). The effects of providing peer feedback on learners' genre awareness in English as a foreign language business letter writing. *Frontiers in Psychology, 13*, 1059555. https://doi.org/10.3389/fpsyg.2022.1059555

Wu, Y., & Schunn C. D. (2021). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal*, *58*(3), 492–526. https://doi.org/10.3102/0002831220945266

Wu, Y., & Schunn, C. D. (2023). Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *Journal of Educational Psychology, 115*(1), 118–142. https://doi.org/10.1037/edu0000768

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., Wang, X., Boud, D., & Lao, H. (2023). The effect of self-assessment on academic performance and the role of explicitness: A meta-analysis. *Assessment & Evaluation in Higher Education, 48*(1), 1–15. https://doi.org/10.1080/02602938.2021.2012644

Zhang, Y., & Schunn, C. D. (2023). Self-regulation of peer feedback quality aspects through different dimensions of experience within prior peer feedback assignments. *Contemporary Educational Psychology, 74*, 102210. https://doi.org/10.1016/j.cedpsych.2023.102210

Zhang, Y., Schunn, C. D., & Wu, Y. (2024). What does it mean to be good at peer reviewing? A multidimensional scaling and cluster analysis study of behavioral indicators of peer feedback literacy. *International Journal of Educational Technology in Higher Education, 21*, 26. https://doi.org/10.1186/s41239-024-00458-1

Zong, Z., Schunn, C. D., & Wang, Y. (2021). Learning to improve the quality peer feedback through experience with peer feedback. *Assessment & Evaluation in Higher Education, 46*(6), 973–992. https://doi.org/10.1080/02602938.2020.1833179

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.