

## Enhancing Peer Feedback Practices With Generative Ai

Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning - CSCL 2025

Greisel, M.; Hornstein, J.; Kollar, I.; Noroozi, O.; Haddadian, G. et al

<https://doi.org/10.22318/cscl2025.921873>

This publication is made publicly available in the institutional repository of Wageningen University and Research, under the terms of article 25fa of the Dutch Copyright Act, also known as the Amendment Taverne.

Article 25fa states that the author of a short scientific work funded either wholly or partially by Dutch public funds is entitled to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed using the principles as determined in the Association of Universities in the Netherlands (VSNU) 'Article 25fa implementation' project. According to these principles research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and / or copyright owner(s) of this work. Any use of the publication or parts of it other than authorised under article 25fa of the Dutch Copyright act is prohibited. Wageningen University & Research and the author(s) of this publication shall not be held responsible or liable for any damages resulting from your (re)use of this publication.

For questions regarding the public availability of this publication please contact [openaccess.library@wur.nl](mailto:openaccess.library@wur.nl)

## Enhancing Peer Feedback Practices with Generative AI

Martin Greisel (chair), University of Augsburg, martin.greisel@uni-a.de  
Julia Hornstein, University of Augsburg, julia.hornstein@uni-a.de  
Ingo Kollar, University of Augsburg, ingo.kollar@uni-a.de  
Omid Noroozi, Wageningen University and Research, omid.noroozi@wur.nl  
Golnoush Haddadian, Georgia State University, ghaddadian1@gsu.edu  
Xingshi Gao, Wageningen University and Research, xingshi.gao@wur.nl  
Maryam Alqassab, Open University of the Netherlands, maryam.alqassab@ou.nl  
Kazem Banihashem, Open University of the Netherlands, kazem.banihashem@ou.nl  
Hassan Khosravi, The University of Queensland, h.khosravi@uq.edu.au  
Stanislav Pozdniakov, The University of Queensland, s.pozdniakov@uq.edu.au  
Christian D. Schunn, University of Pittsburgh, schunn@pitt.edu  
Qiuchen Yu, Central China Normal University, yqc@mails.ccnu.edu.cn

Nikol Rummel (discussant), Ruhr University Bochum, nikol.rummel@rub.de

**Abstract:** Providing peer feedback is a powerful learning activity. However, its potential is not always realized. To improve feedback provision and uptake, AI can augment peer feedback provision and reception processes. However, little is known about how students use AI in the context of peer feedback, what aspects of AI feedback are associated with higher quality of human peer feedback, and how AI feedback can be implemented as part of the instructional design to best facilitate learning. In addition, research needs to find ways to harness the power of AI while minimizing its high resource requirements. This symposium will answer these questions and provide insights into the latest developments in using AI in peer feedback research.

### Generative AI as a potential solution to challenges inherent to peer feedback

Providing feedback is essential for enhancing student learning and performance (Hattie & Timperley, 2007). However, due to heavy workloads, time constraints, and large class sizes, teachers often face challenges in providing feedback. Peer feedback is increasingly regarded as a valuable alternative to teacher feedback (Valero Haro et al., 2023). By actively involving students in the feedback exchange process to critically evaluate their peers' written work, peer feedback not only helps reduce teacher workload but also greatly enhances their learning by activating deep learning processes such as comparing, evaluating, reflecting, and critical reasoning (Kollar & Fischer, 2010; Noroozi et al., 2023). Meta-analyses confirm this learning advantage (Double et al., 2020).

Despite these benefits, students often have problems regarding how to provide high-quality feedback and how to use feedback from their peers to improve their performance. Providing high-quality feedback is a complex task that requires considerable cognitive processing to meticulously evaluate peers' assignments, identify problems, and propose constructive remedies (Banihashem et al., 2024). Consequently, though these complex cognitive processes are key to the undoubted general efficacy of peer feedback, it is not surprising that this feedback often is not of sufficient quality, for example, lacking relevant criticism (Patchan et al., 2013). In terms of feedback reception and uptake, challenges arise from psychological, emotional, and social barriers, including students' perceptions of their peers and the feedback they receive, concerns about trustworthiness, and varying levels of tolerance for receiving and accepting critical feedback (Kerman et al., 2024). These challenges can hinder the potential impact of peer feedback and ultimately affect learning outcomes in a negative way.

Thus, effective interventions are needed to address these challenges. Traditionally, scaffolding has been used to improve the quality of feedback (Hornstein et al., in press). For example, sentence starters guide students to focus their feedback on the most relevant aspects of a task solution. However, such scaffolds are typically not adaptive and thus squander further learning (Sharma et al., 2024). In contrast, generative artificial intelligence (AI) can efficiently analyze written task solutions and produce tailored, individualized feedback (Escalante et al., 2023). Unlike human teachers, generative AI is not overwhelmed by the large amount of text that students typically produce in a peer feedback environment. Instead, AI can use this rich textual data to deliver interventions that are tailored to the individual needs of each student. However, educators need to be careful that AI does not replace learning activities such as providing feedback, thereby eliminating learning opportunities. It is therefore important to find ways of combining AI and peer feedback so that AI enhances (rather than undermines) learning.

## Different ways of how to integrate generative AI into peer feedback

This symposium brings together four empirical papers from different parts of the world (Netherlands, Australia, Germany, USA) that explore how generative AI can address these challenges and improve peer feedback practices in educational settings. Specifically, the papers explore how AI can support the feedback process by improving the quality of feedback provided, fostering deeper cognitive engagement, and scaling feedback activities in ways not previously possible.

*Paper 1* from Noroozi, Haddadian, Gao, Schunn, Alqassab, and Banihashem investigates how many students use generative AI, which AI system specifically they use, and how this use relates to the resulting peer feedback features and the uptake of peer feedback comments. In doing so, the study acknowledges that AI is already part of students' regular learning practices and promotes its competent use by providing instructional support on effectively prompting large language models.

*Paper 2* from Khosravi, Pozdniakov, and Noroozi goes a step further and includes generative AI into the instructional design so that it is readily available to all students. This integration enables the use of the log data that students produce while interacting with the system. This allows the authors to analyze not only the human peer feedback features, but also the AI-generated feedback and how students interact with it.

*Paper 3* from Greisel, Hornstein, and Kollar compares the effects of two different ways of delivering AI feedback. On the one hand, the AI provides feedback on the peers' initial solutions available to the peer reviewers while generating their own feedback. On the other hand, AI provides feedback on the peer reviewers' feedback message. Using a randomized controlled experimental design, the authors show that augmenting human with AI feedback does not necessarily surpass traditional peer feedback. However, AI seems to be able to compete with human peer feedback though replacing the feedback providing process.

*Paper 4* from Schunn and Yu emphasizes efficiency, as the use of large language models is resource-intensive. It describes how to identify feedback quality indicators that can then be applied to peer feedback data without the high cost of LLMs, thus solving both the problem of how to afford to apply LLMs to very large datasets, and how to avoid ethical and privacy concerns. In addition, Paper 4 also demonstrates that AI can help researchers distinguish effective from ineffective instructional scaffolds without having to conduct extensive intervention studies. In this study, the authors use AI to analyze the instructions themselves, rather than student responses. As a result, instructors can learn how to design their scaffolds to improve peer feedback quality.

To summarize, the central question of this symposium examines the extent to which AI-supported scaffolding can improve the effectiveness of peer feedback processes. Paper 1 showed that students already independently use AI to support their peer feedback process. However, this self-initiative is significantly correlated with low self-efficacy beliefs and is primarily limited to the provision of feedback. Paper 2's hypothesis that optimized accessibility through full system integration leads to a significant increase in learning outcomes could not be confirmed in general terms. Rather, classical instructional design elements that influence learning motivation seemed to be decisive. This finding is supported by the results of Study 3, which showed no differential benefit of different AI implementations for students so far. Instead, Study 4 points to how instructionally relevant features can be identified by the help of AI to inform future instructional design and peer feedback scaffolds.

## Paper 1: Supporting peer feedback provision and uptake with genAI

Omid Noroozi, Golnoush Haddadian, Xingshi Gao, Christian D. Schunn, Maryam Alqassab, Seyyed Kazem Banihashem

### Introduction

Peer feedback is acknowledged as a crucial learning strategy especially for large size classes where instructors face limited time recourse and high workload to provide timely and personalized feedback (Valero Haro et al., 2023). Peer feedback is a collaborative strategy where students are actively engaged in critically evaluating peers' work to identify issues and suggest improvements (Kollar & Fischer, 2010; Noroozi et al., 2023). Peer feedback enhances engagement, motivation, satisfaction, and critical skills (Lahza et al., 2025). However, providing effective peer feedback is not an easy task as it requires both subject knowledge and strong feedback skills (Banihashem et al., 2024). Peers' uptake of feedback can be challenging, too. For instance, if students doubt their peers' competence, this perception may discourage them from accepting peer feedback. Additionally, overly critical peer feedback can create emotional tension, making it harder to accept and apply (Kerman et al., 2024). To address these challenges, various frameworks and models (e.g., Lipnevich & Panadero, 2021; Wu & Schunn, 2023) have been proposed to support student in peer feedback, guide their processing and integration of peer feedback, and enhance their peer feedback skills.

While we acknowledge the above-mentioned efforts, advancements in Generative Artificial Intelligence (GenAI) have presented completely new yet promising avenues to potentially elevate peer feedback practices

through the facilitation of scalable, actionable, timely, and personalized feedback. Nevertheless, the empirical research on GenAI's potential to improve peer feedback practices is limited. Exploring how GenAI can be effectively integrated to improve and scale peer feedback practices is critical. It is crucial to understand whether and how students use GenAI tools to enhance their peer feedback practices. Equally important is exploring whether students perceive GenAI as a valuable source of feedback and how they integrate it into the peer feedback process. This study aims to explore students' perceptions and experiences with using GenAI for peer feedback activities, examining the differences in feedback quality between students who utilized GenAI and those who did not. Furthermore, the aim is to investigate how the uptake of peer feedback received varies between students who utilized GenAI and those who did not.

## Method

This study, conducted at a Dutch university, involved 54 graduate students who wrote an argumentative essay, provided feedback to peers, and then revised their own essays based on the feedback received. We analyzed 100 peer feedback sets from 50 students who completed a self-reported post-test survey on the use of GenAI tools. In the beginning of the course, all students received detailed instructions on using GenAI for argumentative essay writing and peer feedback. This included an explanation of the host university's policy on the responsible, ethical, and transparent use of GenAI, along with examples of prompts for leveraging GenAI in argumentative essay writing (e.g., *Revising/refining arguments in favor and against your position on the topic of the essay*; Example: "What are the other possible counter-arguments related to my position on the issue at stake that is not reflected in my essay?"), revising essays based on the received feedback (*Checking if peer feedback is correctly implemented*; Example: "Can you help me check if the new evidence I added effectively strengthens my argument in response to my peer's feedback?"), giving peer feedback (*Suggesting ideas to make your feedback constructive*; Example: "My peer's argument is strong, but it feels like something is missing. What specific suggestions can I offer to enhance their argument?"), and reflecting on received feedback (*Interpreting/explaining unclear parts of feedback*; Example: "What does my peer likely mean when they say this part of my essay needs to be 'more specific and focused'?").

Using event sampling, we segmented each feedback set into individual comments for coding. To measure feedback quality, we used an adjusted coding scheme of feedback quality developed by Wu and Schunn (2020a, 2020b), including those related to non-implementable comments (i.e., no revision-oriented information) and implementable features (cognitive, metacognitive, affective). Specifically, we coded each feedback comment for eight features across three dimensions: cognitive features (identification, explanation, suggestion, and solution), affective features (hedge (problem), hedge (solution), and mitigating praise), and metacognitive features. We also coded each comment based on the high- (e.g. comments with regard to the structure, arguments, and justification) and low-level (e.g. comments with regard to language and flow) issues addressed by the comment. We adapted the social modes of co-construction in collaborative learning (Weinberger & Fischer, 2006) to categorize students' peer feedback uptake behaviors into four types: accept, elaborate, modify, and no uptake. Feedback was coded as "accept" when a specific feedback suggestion was accepted without additional information. Feedback was coded as "elaborate" if feedback was incorporated and expanded, for example, adding some details, examples, or evidence. Feedback was coded as "modify" when a student addressed the problem identified in the comment but did so in a fundamentally different way than the comment suggested. Feedback was coded as "ignore" when a student made no changes related to the suggested feedback. The authors also developed a self-reported post-test survey to measure students' perceptions and experiences using GenAI tools in peer feedback practices. This survey includes (a) tool usage (*whether students used GenAI tools and which tools they used*), and (b) usefulness (*ratings of how helpful GenAI tools were for various essay-writing and peer feedback tasks*).

## Results

The results of the self-reported post-test survey showed that in total, 72% of students used GenAI tools in some capacity, whether for essay writing, feedback provision, or uptake. The remaining 28% did not use GenAI at any stage, as they felt confident in succeeding independently and believed they would learn more by completing the work on their own. Students were more likely to utilize GenAI during the feedback provision process than during the uptake process,  $t = 2.45$ ,  $p < .05$ . In terms of specific GenAI tools used, 86% of students reported using ChatGPT ( $N = 31$ ), 50% used Grammarly ( $N = 18$ ), and 22% used QuillBot ( $N = 8$ ). Additionally, 5% mentioned using Google Bard, Gemini, or AI Checker online ( $N = 2$ ).

Students who reported using GenAI during peer feedback provision provided significantly more suggestions for high-level issues,  $U = 199.5$ ,  $p < .05$ , and significantly less mitigating praise for low-level issues,

$U = 233, p < 0.05$ . When considering whether students employed specific functions (i.e., identifying the issues, explaining the criteria, giving constructive suggestions, and balancing the tone) in the feedback provision process, there was only one significant difference in the corresponding feedback features they provided. When students addressed high level issues, using the providing suggestion function led to more suggestions,  $U = 105.5, p < .05$ . While there were slightly more elaborations and fewer no-uptake behaviors for high-level issues, no significant differences emerged between students who reported using GenAI and those who did not. Thus, according to this observation, the reported use of GenAI was not substantially related to students' uptake behaviors.

## Conclusion and implications

The study reveals mixed use of GenAI tools among students, with less than a third choosing not to use any GenAI tools, primarily due to confidence in their own abilities and a preference for independent learning. Most students who engaged with GenAI did so primarily during the feedback provision process rather than uptake. However, while GenAI use enhanced feedback quality, there was no difference between the group who used GenAI and those who did not in terms of feedback uptake behaviors. Future research is needed to explore how GenAI could be optimally utilized to enhance peer feedback practices, particularly regarding feedback uptake.

## Paper 2: Generative AI feedback to elevate the quality of peer feedback: Empirical evidence for pedagogical design

Hassan Khosravi, Stanislav Pozdniakov, Omid Noroozi

### Problem statement

Peer feedback offers numerous benefits in educational settings, fostering collaboration, critical thinking, and skill development and is particularly beneficial for learning as it activates key cognitive processes such as analysis, evaluation, and reflection during the act of providing feedback. However, challenges persist, particularly in students' ability to provide constructive and actionable feedback due to skill gaps. Generative AI has the potential to address these challenges by reviewing peer feedback in real-time and providing targeted suggestions for improvement. Despite its increasing use in education, there remains limited understanding of the optimal conditions for its use and a lack of robust empirical evidence on the effectiveness of generative AI.

### Aim and research questions

This study addresses this gap by presenting findings from RiPPLE (Khosravi et al., 2019), a platform where students create bite-sized learning resources that are evaluated by a peer review process. During the peer review process, students receive immediate AI-generated feedback on their drafted feedback before submitting it. This AI feedback includes a summary of strengths and actionable suggestions for improvement, helping enhance the quality of peer evaluations. We assess the impact and quality of GenAI feedback on peer feedback from multiple perspectives. Our research is guided by the following questions (1) What are the key characteristics of GenAI feedback on peer feedback? (2) How do students engage with and utilize the AI-generated feedback during the peer feedback process? And (3) How do students perceive the usefulness and effectiveness of the AI feedback?

### Method

To address our research questions, we use a comprehensive data analysis approach, employing both quantitative and qualitative measures. For *RQ1*, we assess the key attributes of GenAI feedback based on 7670 instances of AI feedback on peer reviews. We focus on depth, tone, and scope. Depth is measured by analyzing the length of the feedback text. Tone is evaluated using sentiment analysis, which classifies feedback as positive, negative, or neutral. Scope is examined by categorizing the types of suggestions, such as content-specific guidance or general writing advice. For *RQ2*, we explore how students engage with and utilize AI feedback of the 7670 instances of moderation sessions by analyzing detailed interaction log data. This data captures student actions after receiving AI feedback, including whether they made revisions or chose to submit the original peer feedback without modifications. For *RQ3*, we investigate student perceptions of the quality of AI feedback using 63 instances of optional student ratings on the AI feedback.

### Results

In terms of length, the AI-generated feedback had a *mean word count* of 301.95 words ( $SD = 179.04, IQR = 231.5$ ). The *positive appraisals section* averaged 131.95 words ( $SD = 77.49, IQR = 102.0$ ), while the *suggestions section* averaged 142.21 words ( $SD = 90.63, IQR = 117.0$ ). Feedback length was fairly standard and showed consistent patterns across all cases. To assess tone, we used the Vader sentiment analysis tool, which provides

scores ranging from  $-1$  (most negative) to  $+1$  (most positive). Overall, the AI-generated feedback demonstrated a *highly positive tone*, with a *mean score* of  $0.97$  ( $SD = 0.1$ ,  $IQR = 0.01$ ). The *positive appraisals section* had a *mean score* of  $0.93$  ( $SD = 0.14$ ,  $IQR = 0.05$ ), reflecting a consistently positive sentiment. Similarly, the *suggestions section* exhibited a *mean score* of  $0.89$  ( $SD = 0.2$ ,  $IQR = 0.08$ ), indicating a slightly less positive but constructive tone. These results highlight the overall positivity and balance of the feedback, with variations in tone across different sections. The scope of AI-generated feedback emphasizing strengths of students' moderations is characterized by instances focused on pedagogy ( $N = 11,816$ ,  $78.9\%$ ), highlighting the clarity and constructiveness of students' moderations, followed by disciplinary content ( $N = 2,338$ ,  $15.6\%$ ) and language and writing suggestions ( $N = 830$ ,  $5.5\%$ ). Similarly, most AI-suggestions contained pedagogic guidance ( $N = 8,366$ ,  $57.4\%$ ), suggesting actionable improvements, and writing guidance ( $N = 5,812$ ,  $39.9\%$ ). Subject-specific accuracy was present in the least AI feedback instances ( $N = 390$ ,  $2.7\%$ ). The lower proportion of suggestions containing disciplinary content may reflect the AI's focus on structural or pedagogical improvements over subject-specific corrections, which require deeper domain knowledge and may be less frequently identified as problematic.

In terms of feedback uptake, only  $9\%$  of cases showed students making revisions based on the provided AI feedback, while  $91\%$  opted to submit their peer review without making any changes. This low uptake can be attributed to two key factors: Firstly, in this course, students are primarily assessed on their effort in completing the peer review, which may reduce their motivation to make refinements. Secondly, the current UI design may not be well-optimized for acting on the feedback, and streamlining the workflow process to make it more intuitive could encourage students to engage with and implement the feedback more readily. In terms of student perceptions,  $39$  students ( $62\%$ ) rated it as positive (ratings of  $4$  or  $5$ ),  $11$  students ( $17\%$ ) rated it as neutral (rating of  $3$ ), and  $13$  students ( $21\%$ ) rated it as negative (ratings of  $1$  or  $2$ ). This distribution highlights that the majority of students perceived the feedback positively, while a smaller proportion viewed it as neutral or negative.

## Conclusion and implications

In summary the analysis of AI-generated feedback highlights its generally positive tone and constructive nature, with balanced content across positive appraisals and suggestions, emphasizing pedagogical clarity and actionable improvements over subject-specific corrections. However, feedback uptake by students was limited, suggesting potential barriers related to the course's assessment structure and the usability of the feedback interface. While the majority of students perceived the feedback positively, a smaller group expressed neutral or negative views. These findings point to opportunities for refining the feedback process and interface to enhance usability, encourage engagement, and maximize its impact on student learning.

It is important to also reflect on broader educational, technological, and social implications of incorporating Generative AI into peer feedback systems. *Educationally*, the integration of Generative AI provides a powerful means for instructors to improve the overall quality and reliability of peer feedback. While Generative AI has the potential to assist in addressing common peer review shortcomings, current models are prone to generating hallucinations or providing feedback that may lack contextual relevance or depth. Hence, active instructor involvement remains essential to review and contextualize AI outputs, ensuring that they align with educational objectives and genuinely enhance student learning. *Technologically*, the use of Generative AI in peer feedback is still developing, and best practices for effective implementation are not yet fully established. A robust framework must be developed to ensure data privacy, security, and ethical considerations are met, alongside ensuring transparency and fairness in feedback generation. Additionally, the operation of these algorithms must be designed to provide unbiased and equitable support for a diverse student population. For these tools to be effective and reliable, both students and instructors must develop AI literacy, gaining an understanding of how Generative AI models work, their benefits, and their limitations. *Socially*, AI-powered feedback tools should reinforce, rather than detract from, the collaborative and dialogic essence of peer learning. While AI can streamline and enrich the feedback process, it is critical to maintain human-centered learning interactions. The strategic and ethical use of Generative AI should aim to complement human judgment, encouraging a culture of constructive and meaningful peer engagement. Thus, incorporating AI in a way that upholds and enhances the human-centric nature of education remains a priority.

## Paper 3: Blending human peer feedback with generative AI feedback to improve skill development

Martin Greisel, Julia Hornstein, Ingo Kollar

### Problem statement

Providing feedback to peers can be beneficial for learning because through generating feedback, learners are likely to deeply elaborate on the learning material (Yu & Schunn, 2023). However, students do not always produce high-quality feedback (e.g., lacking relevant criticism; Patchan et al., 2013).

Generative AI may improve the quality of student feedback and task solutions, but also pose the risk to de-skill the students by offloading learning activities to AI, as AI taking over feedback production eliminates the need for deep elaboration when providing feedback. Therefore, we investigated how AI can *augment* learners' deep processing of the learning material and their peer's initial task performance when providing feedback instead of *substituting* it.

### Augmenting feedback provision

At least two ways to augment feedback provision seem possible: First, the human reviewer and the AI both generate feedback on a peer's initial task performance. Then, the reviewer compares both feedback messages and creates a "best-of" version. Only after this step, the review is delivered to the feedback recipient. This approach combines a) the advantages of providing feedback as a learning activity with b) the potential of AI feedback which likely refers to more relevant aspects of the initial task solution in a more specific and elaborated way (Escalante et al., 2023), and c) requires cognitively engaging comparison processes. The downside is that students might copy and paste parts of the AI feedback into their revised feedback without deeply processing it, resulting in less learning gain.

Second, the peer-feedback process may be designed in a way that the human reviewer generates a peer-feedback message, then the AI provides feedback on this peer-feedback message, and then the reviewer revises their feedback accordingly before they send it to the recipient. This approach keeps the reviewer fully accountable for crafting a good feedback message, thereby probably optimizing the reviewer's learning gain.

To investigate their effectiveness, we contrast these two augmentation alternatives ("best-of" and "feedback-on-feedback") with classical peer-feedback ("human-only") and "AI-only" feedback, in which AI substitutes the feedback provision, and look at the adequacy of the resulting feedback and the subsequent evidence-informed problem-solving of pre-service teachers.

### Method

Pre-service teachers participated in two waves to reach a sample of about 400 students. The first wave's data,  $N = 186$ , is presented in this summary; the second wave is currently ongoing and will be added to the data for the conference.

The learning scenario comprised five phases of about one week each. In the first week, students analyzed three teaching problems depicted in a case vignette of a lesson (395 words). For example, one case vignette described a classroom situation in which students were not talking to each other during collaborative learning. The analysis of each problem had to be based on short theory summaries (e.g., cognitive load theory, Sweller, 2005, and ICAP-model, Chi & Wylie, 2014). In Week 2, students provided feedback to two randomly assigned and anonymous peers with the help of guiding questions and a worked example to scaffold feedback production. In Week 3, they revised their feedback. In Week 4, students received their peers' feedback and revised their initial problem analyses. In Week 5, they analyzed a second case vignette.

In a 1x4 experimental design, students were randomly assigned to the four conditions during Week 2 and 3: Students compared their own feedback with AI feedback on their peers' initial task solutions ("best-of-AI and-peer-feedback"), students received AI feedback on their feedback draft and used that to revise it ("AI-feedback-on-peer-feedback"), students revised their feedback without extra guidance ("human-only"), and students only received AI feedback on their own initial task performance ("AI-only"). To generate the AI feedback, we used OpenAI's model GPT-4o-2024-05-13 with prompts that included the material that students had available—task instructions, case vignette, and students' task solutions (and, for the feedback-on-feedback-condition, students' feedback draft).

*Perceived Feedback Adequacy* was assessed with nine items on a Likert-Scale from 1 = *totally disagree* to 9 = *totally agree* comprising the dimensions fairness, utility, and acceptance, which form a global adequacy score (Feedback Perceptions Questionnaire, Strijbos et al., 2021) ( $\omega_h = .82/.85$ ).

*Performance.* Two trained raters coded (Gwet's AC2 = .84) students' problem analyses from a separate case vignette in the post-test regarding their evidence-informed reasoning quality based on a step-wise model (Greisel et al., 2022) from 0 = *severe mistakes/reasoning steps completely missing* to 3 = *all reasoning steps precisely elaborated and correct*.

## Results

We calculated two one factorial ANOVAs with the data of Wave 1. First, the experimental conditions did not differ in regard to the adequacy of the feedback in the receivers' eyes,  $F(3, 145) = 0.24, p = .868, \eta^2 = 0.005$ . Second, the evidence-informed reasoning performance in the post-test did not differ between conditions,  $F(3, 144) = 1.43, p = .238, \eta^2 = 0.03$ .

## Conclusion and implications

The results seem to imply that the quality of AI feedback reaches quality levels comparable to human student feedback. Its quality even might have been able to compensate for the eliminated learning activity of feedback providing. However, future research needs to address how AI may be combined with human activities to raise the quality of the feedback and the learning gains while providing it beyond human-only peer-feedback. Practitioners might take away that even zero-shot, out-of-the-box large language models can generate valuable feedback today.

## Paper 4: Wisely guiding rather than unwisely automating the overwhelming task of analyzing peer feedback data

Christian D. Schunn and Qiuchen Yu

### Problem statement

Peer feedback has clear benefits for learning based upon the cognitive processes it inherently involves knowledge construction (e.g., as part of giving explanations or advice for specific improvements; Wu & Schunn, 2023). As peer feedback continues to grow in popularity, researchers have the opportunity to analyze increasingly larger datasets. On the plus side, this enables researchers to entertain for-whom and under-what-circumstances questions. Such questions are particularly important since the 9+ meta-analyses on peer feedback/assessment published in the last 10 years all find large, unexplained heterogeneity in effects (e.g., Double et al., 2020). On the minus side, it is challenging to analyze all the data. Researchers taking on larger data have tended to rely on simple quantitative measures like the ratings produced by students or comment counts/word counts on comments produced (e.g., Yu & Schunn, 2023). However, this limits potential research questions and misses important variation in learning opportunities. For example, existing research has established that a number of richer features of comments matter for learning within the comment provider and comment receiver. For example, one study found that comment receivers are more likely to learn when they receive comments containing explanations, and comment providers are more likely to learn when the provided comments contain constructive advice and explanations (Wu & Schunn, 2023). Note that this finding was based upon data from approximately 370 students, and it involved over a year of laborious coding of peer feedback comments by hand. A later study by another group found no benefit of explanations in a very different context, again after a very extensive hand-coding effort (He & Gao, 2023). Efficient methods are needed to identify general patterns and tease apart moderation by task, context, scaffolds, and learner characteristics.

AI has long offered the possibility of automating coding of complex text data (including peer feedback data; Xiong, et al., 2012), and the rise of generative AI has increased this potential analysis pathway. A number of scholars have discussed strategies for applying generative AI to develop codebooks (that then are used by humans) as well as using generative AI to apply given codebooks to large datasets (e.g., Liu et al., 2024; Barany et al., 2024). It is becoming clear that there are a lot of devils in the details regarding when acceptable results are obtained (e.g., which kinds of coding tasks and which kinds of prompts), and generative AI should definitely NOT be applied in a fully automated way to the task given the relatively high risk for hallucinations, omissions, and odd redundancies that often occur in outputs. There are also ethical concerns (can AI have access to student data?) and financial concerns (particularly when applied to very large datasets).

### Case 1

Here we present two uses of generative AI to guide data analysis but not automate it when applied to very large peer feedback datasets. In the first use, we asked ChatGPT4 to code the contents of a sample peer feedback dataset in terms of the presence of explanations and suggestions in comments. We then asked it to describe how it did so, and it described a keyword search method. We examined the suggested keywords/key phrases, making some

adaptations (e.g., adding additional ones, removing superficial ones) based upon our past experiences in coding peer feedback comments. We were initially suspicious of the accuracy of this simple approach since prior attempts to automate feature recognition required using more complex decision trees involving keywords and other text features (Xiong et al., 2012). However, given the more robust set of keywords (29 for solutions; 24 for explanations), we thought it might work. We implemented the keyword search in MS Excel and applied it to a very large dataset involving over 625,000 comments from hundreds of courses, thereby addressing ethical and financial concerns. We compared codes against human coders for 300 randomly selected comments, finding very high reliability (Kappa = 0.91 for suggestion, Kappa = 0.91 for explanation). The resulting coded dataset was then used in one publication (Yu & Schunn, in press) and in several ongoing projects.

## Case 2

In the second use, we built upon a prior meta-regression study applied to a dataset of 243 courses involving multiple consecutive peer feedback assignments (Yu & Schunn, 2023). The peer feedback experiences in one assignment were statistically related to growth in task performance in the next assignment using regression. Meta-regression was then applied to examine the consistency of statistical relationships. The study found that the length of comments provided was the strongest predictor of task performance growth and that the length of comments received was not related to task performance growth, at least on average. However, across the more than 500 assignments included in the dataset, there was very large heterogeneity in the relationships that was not noise and not explainable by simple contextual factors like the size of the course, the discipline, the level of the students, etc. Here, we followed up on that heterogeneity and used ChatGPT to suggest possible differences between those assignments showing large effects of providing feedback and those assignments showing smaller effects of providing feedback (and similarly for variation in receiving feedback effects), considering the nature of the feedback scaffolds included in the various assignments. In other words, ChatGPT took the commenting scaffolds and rating rubrics as data, rather than student comments as data, addressing both ethical and financial concerns. It made some suggestions regarding possible differences (for example emphasizing basic understanding vs. complex analysis; emphasizing compliance/correctness vs. overall quality and integration of ideas), which we formalized into coding schemes that we applied by hand to the comments and rubrics data.

## Conclusion and implications

Beyond providing a model for how AI can be used to improve data analysis of massive peer feedback datasets, the specific applications produced results that can be directly integrated into peer feedback scaffolds. For example, the automated methods for detecting features can be used to efficiently give direct guidance to students on the quality of their feedback or reports to teachers about where additional interventions might be needed (e.g., students, in general, are lacking explanations in this assignment or on this dimension; these specific students generally are lacking in giving concrete suggestions for improvement; these specific students received too few suggestions in their received comments; etc.). In addition, the AI-provided ideas for new dimensions of scaffolds could influence the effectiveness of peer feedback and can be the subject of new experiments as well as be part of new teacher-facing AI tools (e.g., new wizards that help teachers set up more effective peer feedback assignments).

## References

- Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., ... & Baker, R. S. (2024, July). ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In *International Conference on Artificial Intelligence in Education* (pp. 134-149). Cham: Springer Nature Switzerland.
- Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, 32(2), 481–509.
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Greisel, M., Wekerle, C., Wilkes, T., Stark, R., & Kollar, I. (2022). Pre-service teachers' evidence-informed reasoning: Do attitudes, subjective norms, and self-efficacy facilitate the use of scientific theories to analyze teaching problems? *Psychology Learning & Teaching*, 22(1), 20–38. <https://doi.org/10.1177/14757257221113942>

- He, W., & Gao, Y. (2023). Explicating peer feedback quality and its impact on feedback implementation in EFL writing. *Frontiers in Psychology, 14*, 1177094.
- Hornstein, J., Keller, M. V., Greisel, M., Dresel, M., & Kollar, I. (in press). Enhancing the peer-feedback process through instructional support: A meta-analysis. *Educational Psychology Review*.
- Kerman, N. T., Noroozi, O., Banihashem, S. K., Karami, M., & Biemans, H. J. (2024). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments, 32*(2), 614-626. <https://doi.org/10.1080/10494820.2022.2093914>
- Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction, 20*(4), 344-348. <https://doi.org/10.1016/j.learninstruc.2009.08.005>
- Lahza, H. F., Demartini, G., Noroozi, O., Gašević, D., Sadiq, S., & Khosravi, H. (2025). Enhancing peer feedback provision through user interface scaffolding: A comparative examination of scripting and self-monitoring techniques. *Computers & Education, 230*, 105260. <https://doi.org/10.1016/j.compedu.2025.105260>
- Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*, 720195.
- Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R. S. (2024, November). Assessing the potential and limits of large language models in qualitative coding. In *International Conference on Quantitative Ethnography* (pp. 89-103). Cham: Springer Nature Switzerland.
- Noroozi, O., Banihashem, S. K., Biemans, H. J., Smits, M., Vervoort, M. T., & Verbaan, C. L. (2023). Design, implementation, and evaluation of an online supported peer feedback module to enhance students' argumentative essay quality. *Education and Information Technologies, 28*(10), 12757-12784. <https://doi.org/10.1007/s10639-023-11683-y>
- Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*(2), 381-405. <https://doi.org/10.1007/s11251-012-9236-3>
- Sharma, K., Nguyen, A., & Hong, Y. (2024). Self-regulation and shared regulation in collaborative learning in adaptive digital learning environments: A systematic review of empirical studies. *British Journal of Educational Technology, 55*(4), 1398-1436. <https://doi.org/10.1111/bjet.13459>
- Strijbos, J.-W., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the Feedback Perceptions Questionnaire. *Studies in Educational Evaluation, 68*, 100980. <https://doi.org/10.1016/j.stueduc.2021.100980>
- Valero Haro, A., Noroozi, O., Biemans, H. J., Mulder, M., & Banihashem, S. K. (2023). How does the type of online peer feedback influence feedback quality, argumentative essay writing quality, and domain-specific learning? *Interactive Learning Environments, 1-20*. <https://doi.org/10.1080/10494820.2023.2215822>
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71-95. <https://doi.org/https://doi.org/10.1016/j.compedu.2005.04.003>
- Wu, Y., & Schunn, C. D. (2020a). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 101826. <https://doi.org/https://doi.org/10.1016/j.cedpsych.2019.101826>
- Wu, Y., & Schunn, C. D. (2020b). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology, 62*, 101897. <https://doi.org/https://doi.org/10.1016/j.cedpsych.2020.101897>
- Wu, Y. & Schunn, C. D. (2023). Passive, active, and constructive engagement with peer feedback A revised model of learning from peer feedback. *Contemporary Educational Psychology, 73*, 102160.
- Xiong, W., Litman, D., & Schunn, C. (2012). Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research, 4*(2), 155-176.
- Yu, Q., & Schunn, C. D. (2023). Understanding the what and when of peer feedback benefits for performance and transfer. *Computers in Human Behavior, 147*, 107857. <https://doi.org/10.1016/j.chb.2023.107857>
- Yu, Q. & Schunn, C. D. (in press). Which experiences are consistently associated with other-regulation of peer feedback length? *Journal of Educational Psychology*.

## Acknowledgments

Paper 3 was supported by the Stiftung Innovation in der Hochschullehre; FBM-2020.