Promoting pre-service teachers' evidence-informed reasoning through peer-feedback: effects of a feedback provision scaffold and a feedback integration scaffold

Julia Hornstein, Martin Greisel, Ingo Kollar

Angaben zur Veröffentlichung / Publication details:

Hornstein, Julia, Martin Greisel, and Ingo Kollar. 2025. "Promoting pre-service teachers' evidence-informed reasoning through peer-feedback: effects of a feedback provision scaffold and a feedback integration scaffold." Instructional Science 53 (6): 1429–55. https://doi.org/10.1007/s11251-025-09745-4.

Instructional Science (2025) 53:1429–1455
https://doi.org/10.1007/s11251-025-09745-4

ORIGINAL RESEARCH

Promoting Pre-Service Teachers' Evidence-Informed Reasoning Through Peer-Feedback: Effects of a Feedback Provision Scaffold and a Feedback Integration Scaffold

Julia Hornstein[1] · **Martin Greisel**[1] · **Ingo Kollar**[1]

Received: 12 July 2024 / Accepted: 12 August 2025 / Published online: 2 September 2025

Abstract
Peer-feedback can be an effective way to improve pre-service teachers' evidence-informed reasoning skills (=EIR skills). However, because learners often struggle to provide high-quality feedback or to make sense of the feedback they receive, they need scaffolding. In this study, $N=254$ pre-service teachers analyzed an authentic classroom case from the perspective of two educational theories and related evidence, provided feedback to two peers, and processed two feedback messages they received to revise their initial case analyses. In a 2×2 factorial between-subjects design, we investigate the effects of a feedback provision scaffold and a feedback integration scaffold, and their interaction on pre-service teachers' feedback quality, their evidence-informed reasoning recognition skills (=EIR recognition skills), and their evidence-informed reasoning application skills (=EIR application skills). The results show that feedback quality was significantly improved by the feedback provision scaffold. For EIR recognition skills, we found a negative interaction effect of the two scaffolds, indicating that they should rather be presented alone instead of in combination with each other. For EIR application skills, we found no significant effects of the two scaffolds. Thus, the study shows that providing low threshold prompts and sample illustrations of good feedback already is sufficient to raise feedback quality greatly. However, more research is needed on how to optimally combine these two scaffolds.

Keywords Peer-feedback · Scaffolding · Evidence-informed reasoning · Higher education · Pre-service teachers

✉ Julia Hornstein
julia.hornstein@uni-a.de

Martin Greisel
martin.greisel@uni-a.de

Ingo Kollar
ingo.kollar@uni-a.de

[1] Department of Educational Psychology, Augsburg, GermanyUniversitätsstraße 10, 86159

Background and Aims of the Study

Peer-feedback is an effective way of supporting student learning (e.g., Double et al., 2020), particularly in written tasks (e.g., Huisman et al., 2019). In pre-service teacher education, such written tasks may involve pre-service teachers analyzing classroom situations. We assume that peer-feedback should also be effective to support pre-service teachers in this process, and more specifically in written evidence-informed reasoning about such situations (e.g., Gorard, 2020). Yet, research on this issue is scarce. In theory, evidence-informed reasoning skills (=EIR skills) are required to analyze classroom situations in accordance with educational evidence instead of based on gut feeling, conventional wisdom, or experiential knowledge. These should be developed during pre-service teacher education (e.g., Voss, 2022). Yet, studies have revealed that pre-service teachers often struggle with evidence-informed reasoning when being confronted with authentic teaching situations (e.g., Rousseau & Gunia, 2016). First, they often refer to non-scientific knowledge, even when they are given access to scientific information (Menz et al., 2021). Second, their reasoning seldom reflects a systematic, hypothesis-driven process (Csanadi et al., 2021).

One possible way to support pre-service teachers' in their evidence-informed reasoning about classroom situations might be through peer-feedback (e.g., Panadero & Alqassab, 2019): By providing feedback on their peers' analyses of authentic teaching situations, and by implementing the feedback they receive from their peers in their own analyses, pre-service teachers can be supported in constructing knowledge about educational theories and how to apply them to educational practice (e.g., Chi & Wylie, 2014; Double et al., 2020; Ion et al., 2018). However, studies show that peer-feedback does not always work well: On the one hand, learners generally often have difficulty formulating high-quality feedback messages (e.g., Alemdag & Yildirim, 2022). On the other hand, once they have received feedback from their peers, learners often struggle to implement the feedback they receive from their peers in their revision, especially when they receive feedback from more than one peer. When receiving feedback from multiple peers, they first need to integrate these messages with each other, especially when these messages differ in their evaluations and suggested solutions (e.g., List & Alexander, 2019). Only after integration, learners can readily implement multiple feedback messages to revise their initial case analyses (e.g., Carless & Boud, 2018; Wichmann et al., 2018). Thus, learners and therefore also pre-service teachers need scaffolding on both how to provide high-quality feedback messages and how to integrate multiple feedback messages with each other to implement this feedback in their own task solutions (e.g., Gielen & De Wever, 2015; Hornstein et al., 2025). The present study therefore investigates the effects of a *feedback provision scaffold* (i.e., guidance on how to provide high-quality feedback messages to their peers' analyses of an authentic classroom case), as well as a *feedback integration scaffold* (i.e., guidance on how to integrate feedback they receive from multiple peers) on (a) the quality of the feedback messages pre-service teachers provide to their peers' analyses of authentic classroom cases, as well as their acquisition of (b) evidence-informed reasoning recognition skills (=EIR recognition skills), and of (c) evidence-informed reasoning application skills (=EIR application skills).

Evidence-Informed Reasoning Skills of Pre-Service Teachers

Every day, teachers face a range of situations in their classrooms that require thoughtful and goal-directed action (e.g., Slavin, 2020). Fortunately, educational research offers abundant evi-

dence teachers can draw on to solve many of these situations. The process of applying related knowledge has been termed "evidence-informed reasoning" (e.g., Kollar et al., 2023). In this study, we focus primarily on problematic teaching situations that may—if teachers do not handle them well—result in severe consequences for their pupils (e.g., not understanding important learning material, collecting demotivating experiences of failure, etc.). When using the term "evidence-informed reasoning", we follow a rather broad concept of "educational evidence" which includes both educational theories and empirical evidence, as theories are typically tested and iteratively refined through empirical evidence (at least in psychologically oriented research; see Kollar et al., 2023). Together with others (e.g., Gorard, 2020; König et al., 2020), we argue that orienting one's own teaching actions on available theoretical and empirical evidence from educational research has a strong potential to cope with teaching situations in an effective and professional way (e.g., Shavelson, 2020).

Based on research on teachers' professional vision (e.g., Gegenfurtner et al., 2020; Seidel & Stürmer, 2014), Greisel et al. (2022) suggested a normative model of EIR skills: According to them, optimal evidence-informed reasoning about classroom situations includes five steps: First, teachers need to identify a problem that requires further attention (*problem identification*). In the second step, the teacher describes what exactly the problem in the current situation is (*problem description*). This description paves the way for the third step, in which the teacher ideally explains the problem in a systematic way, that is, they reconstruct the situation by describing causal relationships under consideration of scientific evidence from educational research (*problem explanation*). In the fourth step, the teacher formulates a desired goal state (*goal setting*), and in the last step, they decide which particular action is likely to reach that goal (*select actions*).

Clearly, pre-service teachers do not acquire these EIR skills automatically (e.g., Ferguson & Bråten, 2022; Wilkes & Stark, 2023). For example, Csanadi et al. (2021) showed that pre-service teachers rarely form hypotheses on what causes the problems they encounter in authentic classroom cases ("problem explanation" in Greisel et al., 2022). Instead, they often generate solutions for the problems without a systematic reasoning process. To avoid such problems, we argue that pre-service teacher educators should design learning opportunities that take skill acquisition theories and related empirical evidence into account (e.g., Anderson & Lebiere, 1998; Wecker & Fischer, 2014).

According to this research, skills develop in a sequence of three phases: At first, pre-service teachers need to acquire declarative knowledge about the to-be-learnt skill. This knowledge should help them reproduce what is important during evidence-informed reasoning about teaching situations and identify instances in which they or (pre- or in-service) teachers they observe apply the skill appropriately (=EIR recognition skills). After that, this declarative knowledge needs to be converted into procedural knowledge (=EIR application skills). For this, practicing is important (Wecker & Fischer, 2014). Finally, the proceduralized knowledge is fine-tuned through its extensive application in practice situations. Since this third step extends way beyond pre-service teacher education, we only focus on the first two phases in this article.

These considerations bear three consequences for pre-service teacher education: First, pre-service teachers should initially be explicitly taught about the steps of evidence-informed reasoning, for example based on the normative model suggested by Greisel et al. (2022). Second, they should receive opportunities to practice the application of evidence-informed reasoning. For this, having them analyze authentic classroom cases bears particular potential (e.g., Helleve et al., 2023). Third, since applying a newly acquired skill to authentic classroom cases typically is difficult (e.g., Bjorklund et al., 1997), pre-service teachers require additional scaffolding during

this proceduralization phase. As indicated above, one way to provide such scaffolding might be to have pre-service teachers provide feedback on each other's case analyses and to integrate that feedback into their analyses (peer-feedback; e.g., Panadero & Alqassab, 2019). Thus, the following section explores the potential benefits of peer-feedback for learning.

Benefits of Peer-Feedback to Support Reasoning and Learning

Feedback is one of the most powerful ways to enhance student learning (e.g., Hattie, 2009). It refers to "all post-response information that is provided to a learner to inform the learner about his or her actual state of learning or performance" (Narciss, 2008, p.127), with the aim to close the gap between what was understood and what should have been understood. Peer-feedback, as a specific form of feedback, is a reciprocal process in which learners evaluate each other's performance (e.g., Baker, 2016; Hattie & Timperley, 2007). It has been shown to be very effective to support reasoning and learning (e.g., Li et al., 2020). According to Bauer et al. (2023), setting up a peer-feedback process includes five steps: First, during the *design phase*, the teacher develops a peer-feedback scenario, which specifies the task and determines which learners provide feedback to whom. In the context of evidence-informed reasoning, a task might be to analyze an authentic classroom case that describes a problematic teaching situation. Second, the learners complete the task (like analyzing a classroom case) during the so-called *task performance phase.* Third, during the *feedback provision phase*, they evaluate the quality of their peers' initial task performance (e.g., their initial case analysis). Fourth, in the *feedback reception phase*, learners process the feedback they received from their peers and revise their task solution (e.g., their case analysis) based on the feedback they received. Fifth, in the *evaluation phase*, they evaluate the peer-feedback process and the learning outcomes.

The potential of peer-feedback to support pre-service teachers' evidence-informed reasoning about authentic classroom cases lies in the active role that learners adopt during this process: When *providing* feedback, learners first read their peers' case analyses, which may provide them with new ideas on how to approach the task and lead to compare them with their own analyses. Also, evaluating their peers' analyses involves evaluation processes, which represents a higher-order cognitive process that includes applying criteria, diagnosing problems, and generating solutions (e.g., Carless & Boud, 2018; van Popta et al., 2017). Furthermore, writing a feedback message implies making one's own knowledge explicit which requires learners to systematize their knowledge to be able to produce a coherent text (e.g., Cho & Cho, 2011; Lundstrom & Baker, 2009). Moreover, when peer-feedback is used to support evidence-informed reasoning, learners practice reasoning based on educational evidence also while providing feedback, not only while producing their initial case analysis (e.g., van Popta et al., 2017).

In turn, when *receiving* feedback, learners ideally learn about what they should have done, how their solutions compare to these task fulfillment criteria, and how they can improve (e.g., Hattie & Timperley, 2007; Nicol et al., 2014). For this to be possible, learners need to actively engage with the feedback they receive. In particular, they need to assess the extent to which the feedback they receive is important and correct, and how they can revise their initial task solution on the basis of this feedback (e.g., Li & Grion, 2019; Strijbos et al., 2010). They do this by comparing the feedback's content with their own task solution, which can help them learn new knowledge or broaden their current understanding (e.g., Nicol & McCallum, 2022).

Finally, the *revision* phase requires that learners not only read the feedback they received, but also engage with, elaborate on, and act on the feedback (Carless & Boud, 2018), which should foster deep learning (e.g., Nicol & McCallum, 2022; Nicol et al., 2014).

Two recent meta-analyses confirm the beneficial effects of peer-feedback: First, Li et al. (2020) examined the general effect of peer-feedback compared to no peer-feedback on performance and showed that peer-feedback can improve performance by 0.291 standard deviations. Second, Double et al. (2020) found a small to medium effect of peer-feedback compared to teacher feedback and conditions without feedback on performance. While both meta-analyses showed that peer-feedback can be effective across a wide range of educational contexts, none of them included primary studies that investigated the effects of peer-feedback to foster evidence-informed reasoning of (pre- or in-service) teachers. Therefore, the present study tries to fill this gap by applying peer-feedback to support pre-service teachers' EIR skills.

Low Feedback Quality and Low Uptake of Feedback as Challenges During Peer-Feedback

Even though peer-feedback has a strong potential to support reasoning and learning, this potential does not always materialize (e.g., Double et al., 2020). On the one hand, learners may lack individual prior experience with peer-feedback and therefore necessary skills and prior knowledge of the subject matter at hand, on the other hand, if the task learners are supposed to work on in a peer-feedback scenario is too complex, this might lead to high intrinsic cognitive load (e.g., Alqassab et al., 2018; Sweller, 1999; van Zundert, Sluijsmans et al., 2012). Both may lead to (1) a low quality of the feedback messages that peers provide to each other (e.g., Alemdag & Yildirim, 2022), and to (2) learners struggling in taking up feedback in a way that would benefit their revisions and ultimately their learning (e.g., Carless & Boud, 2018).

Concerning (1), the literature offers several approaches to conceptualize peer-feedback quality. For Hattie and Timperley (2007), for example, high-quality feedback includes information that reminds the feedback recipient of the criteria to fulfil the task (feed up), comments on their performance (feed back), and offers suggestions for revision (feed forward). Other researchers emphasize that the way the comments are written is another crucial component of high-quality feedback. For example, comments should be specific, illustrative, and explicitly localize the aspects in the case analyses that require the feedback receiver's attention (e.g., Gielen & De Wever, 2015; Patchan et al., 2016). Also, high-quality feedback includes giving reasons why certain elements are right or wrong using examples or explanations (e.g., Gielen & De Wever, 2015). However, research indicates that learners rarely provide their peers with high-quality feedback without guidance (Sluijsmans et al., 2002). One possible way of supporting them could be scaffolds, which we examine in more detail in this study.

With respect to (2), even if peers manage to provide high-quality feedback, the receiver will not automatically benefit from it (e.g., Lui & Andrade, 2022; Wu & Schunn, 2020, 2021). For this, learners need to understand the feedback they receive, process it in a mindful way (e.g., Berndt et al., 2018), and revise their own work based on the feedback (e.g., van der Pol et al., 2008). Walker (2015), however, showed that learners often have problems making use of the feedback they receive: In her study, learners only used about half of the feedback they received from their peers for revision. This may be due to several reasons, such as a low perceived appropriateness of the feedback, negative affect caused by the feedback, or a low willingness to improve at all

(Strijbos et al., 2010). Walker (2015) also found that especially feedback that was vague or grammatically incorrect was hardly taken up in learners' revisions.

Such problems can even be intensified when multiple peer-feedback messages need to be integrated with one another as a prerequisite for a successful implementation, that is when learners receive feedback on their initial task solutions (e.g., case analyses) from more than one peer (e.g., Barzilai et al., 2018; Cho & MacArthur, 2011). When they receive multiple peer-feedback messages, learners need to engage actively with these messages, compare and possibly combine their content as a basis for revision (e.g., Barzilai et al., 2018; Boscolo et al., 2007; List & Alexander, 2019). This integration is challenging, especially when the different peer-feedback messages contain conflicting recommendations (e.g., Hsiao & Rajagopal, 2023). Integration involves several cognitive processes, such as the selection, processing, analysis, and synthesis of information from different sources (e.g., Cerdán & Vidal-Abarca, 2008; Rouet & Britt, 2011). During integration, learners may encounter various difficulties, such as not recognizing inconsistencies between the peer-feedback messages or not having strategies to deal with them effectively (e.g., Bråten et al., 2014; Stadtler et al., 2014). For example, in a study by Jurkowski (2018), the learners wrote a term paper and received multiple peer-feedback comments on their initial draft. Without guidance, learners turned out to implement less correct and more incorrect peer comments—if they implemented the comments into their revision at all. Thus, scaffolding on how to integrate multiple peer-feedback messages seems to be warranted.

Scaffolding Feedback Provision and Feedback Processing

Scaffolding may focus on different parts of the peer-feedback process suggested by Bauer et al. (2023). In this article, we focus on (1) feedback provision and (2) feedback reception. With respect to (1), several studies investigated how learners can be promoted in providing high-quality peer-feedback. One effective scaffold in this respect is prompting. A study by Latifi et al. (2021), for example, showed that learners who either received feed back or feed forward prompts or a combination of both performed significantly better in argumentative peer learning quality, quality of argumentative essay writing, and a domain-specific knowledge test than learners who were not scaffolded during feedback provision (for similar results, see Bürgermeister et al., 2021). Additionally, Gyamfi et al. (2024) demonstrated that presenting learners with guidelines and examples on how to provide feedback, improved the length and level of detail of learners' feedback messages and led to a more frequent occurrence of high-quality feedback features, especially in terms of specificity and constructive suggestions. Yet, whether such prompts and examples also work to help pre-service teachers acquire EIR skills is an open question.

With respect to (2) feedback reception, in the aforementioned study by Jurkowski (2018), half of the learners received a question prompt that asked feedback recipients for content-related arguments for and against implementing the comments they received from their peers, while the other half did not. The results showed that the question prompt helped learners to better identify which comments from the feedback message were correct and which comments were incorrect. However, they still had problems incorporating these comments in their revisions. Interestingly though, there has been little research on how best to promote the integration of *multiple* peer-feedback messages. This integration could be a major barrier to the implementation of feedback, especially in ill-structured tasks where conflicting feedback is likely (e.g., Hsiao & Rajagopal, 2023). Therefore, we take a closer look at how an effective scaffold to support the integration of multiple peer-feedback messages in the context of evidence-informed reasoning might look like.

Even though examining integration processes of multiple documents outside of a peer-feedback context, results of a meta-analysis by Barzilai et al. (2018) indicate that explicit instruction on how learners should integrate different documents with each other can have a positive influence on integration. In addition, Maier and Richter (2014) showed that prompting metacognitive strategies (e.g., asking learners explicitly to filter out congruent and contradictory aspects of the texts) can have a positive impact on integration.

The Current Study

The literature shows that peer-feedback can support learning processes, especially in written tasks (e.g., Huisman et al., 2019). For pre-service teachers to solve problematic teaching situations, an important skill to acquire is evidence-informed reasoning about classroom situations. This is often taught in teacher education having pre-service teachers analyze authentic classroom cases in a written form (e.g., Helleve et al., 2023). However, research shows that pre-service teachers often struggle with producing high-level case analyses (e.g., Wilkes & Stark, 2023). While peer-feedback has the potential to help in this context (e.g., Panadero & Alqassab, 2019), learners often have difficulties in providing and processing feedback at a high level, and therefore require scaffolding, both for feedback provision and feedback integration (e.g., Alemdag & Yildirim, 2022; Carless & Boud, 2018; Hornstein et al., 2025). Such scaffolding may take the form of prompts combined with sample solutions to reduce high intrinsic cognitive load. However, research on the effects of such scaffolds in the context of pre-service teachers' evidence-informed reasoning about classroom situations is lacking. Therefore, this study aims to answer the following research questions:

1. What is the effect of a feedback provision scaffold on the quality of the feedback messages pre-service teachers provide their peers with during evidence-informed reasoning about classroom situations?

   H1: We assume that pre-service teachers who are provided with a feedback provision scaffold formulate higher quality feedback messages than pre-service teachers who are not provided with a feedback provision scaffold.

2. What are the effects of (a) a feedback provision scaffold and (b) a feedback integration scaffold on pre-service teachers' EIR *recognition* skills?

   H2a: We expect that pre-service teachers who are provided with a feedback provision scaffold can distinguish high-quality analyses of teaching problems from low-quality analyses better than pre-service teachers who do not receive a feedback provision scaffold.

   H2b: We expect that pre-service teachers who are provided with a feedback integration scaffold can distinguish high-quality analyses of teaching problems from low-quality analyses better than pre-service teachers who do not receive a feedback integration scaffold.

3. What are the effects of (a) a feedback provision scaffold and (b) a feedback integration scaffold on the acquisition of EIR *application* skills?

H3a: We anticipate that when asked to analyze classroom situations, pre-service teachers who are provided with a feedback provision scaffold apply evidence-informed reasoning at a higher level than pre-service teachers who are not provided with a feedback provision scaffold.

H3b: We anticipate that when asked to analyze teaching problems, pre-service teachers who are provided with a feedback integration scaffold apply evidence-informed reasoning at a higher level than pre-service teachers who are not provided with a feedback integration scaffold.

All hypotheses were pre-registered (https://osf.io/ndjxa)[1]. As we had no specific hyp otheses regarding the interaction of the two scaffolds, we investigated these effects only exploratively.

Method

Sample and Design

We ran an experimental study with $N=254$ pre-service teachers from a German university with a mean age of $M=22.56$ ($SD=4.30$), and who were in their fourth semester on average ($M=4.61$, $SD=1.20$). The percentage of female participants was 77.95%, which is typical for cohorts of pre-service teachers in Germany. At this stage during their studies, the participants have little practical experience of teaching and little experience with peer-feedback. The study was embedded in three parallel, regular courses with the same content, in which participants worked in a web-based learning environment that presented them with authentic case vignettes and with relevant educational evidence. We established a 2×2 factorial randomized between-subjects design with the independent variables "feedback provision scaffold" (available vs. not available) and "feedback integration scaffold" (available vs. not available). Across the three courses, participants were randomly assigned to one of the four experimental conditions. Within each condition and thereby surpassing the original course enrollment, participants were randomly assigned to peers to provide feedback to. For this reason, peers received feedback from different peers than whom they provided feedback to. This peer-feedback process was double-blind. Participation in the peer-feedback exercise was mandatory for receiving course credit. Yet, participants had the opportunity to object against using their data for research purposes, without consequences. In this case, their data were removed from the data set. The study complies with the ethical criteria of the German Psychology Association and European privacy protection regulations. The data are available here: https://osf.io/vsbqd/?view_only=7657506b6a1244cb8e0d562e4e9ce8d0.

[1] In this study, we only examine the hypotheses regarding the effects of the scaffolds for providing and receiving feedback, so the numbering here differs from the numbering in the pre-registration.

Procedure

The complete peer-feedback scenario span over four weeks (see Fig. 1). Participants were told that they could complete all these tasks from home, using the universities' learning management system. The procedure followed the feedback process model proposed by Bauer et al. (2023). After we had specified the task and that feedback pairings would be random within conditions (*design phase*), participants were supposed to write an initial analysis of the classroom vignette (*task performance phase*). During this phase (Week 1), participants were provided with short summaries (795 and 772 words) of two educational theories, namely (a) cognitive load theory (Sweller, 1999) and (b) the ICAP model (Chi & Wylie, 2014), which they were supposed to apply to the problems in the case vignette. Each of these texts included the basic concepts and assumptions of the respective theories as well as short descriptions of one empirical study that supports the specific theory. Participants' task was to individually analyze the written case vignette that described an authentic classroom situation in which a teacher tries to explain the topic 'fairy tales and fables' to her pupils. The case vignette describes that the pupils encounter several problems when working on the topic. For example, they have difficulties understanding the texts, because they have too little prior knowledge about the topic which the teacher does not address adequately (high intrinsic load, CLT). Similarly, the group work the teacher employs mainly evokes passive rather than interactive activities (ICAP). In total, the case vignette included six problems (see Supplementary Information, Appendix A). Participants were asked to identify and analyze each of these problems in an evidence-informed manner, using the five steps "problem identification", "problem description", "problem explanation", "goal setting", and "selecting actions" (Greisel et al., 2022). To illustrate this and to reduce cognitive load regarding evidence-informed reasoning, we gave them an example of the five steps they had access to throughout the peer-feedback process. During the *feedback provision phase* (Week 2), participants subjectively evaluated the quality of the case analyses they received from two randomly assigned peers. For each peer, they then provided feedback on the first three of the six problem analyses (to reduce the workload), either being supported by the feedback provision scaffold or not. During the *feedback reception and revision phase* (Week 3), participants first evaluated the quality of the feedback messages they received from their peers in a questionnaire. Then, they were asked to integrate the two feedback messages and implement their suggestions in the revision of the initial case analysis (either supported by the feedback integration scaffold or not). During Week 4, participants completed a posttest which measured their EIR recognition and application skills utilizing a second case vignette (see below and Supplementary Information, Appendix B for the second case vignette).

Independent Variables

We experimentally varied whether or not participants received a feedback provision scaffold (during Week 2) and/or a feedback integration scaffold (during Week 3). Regarding feedback provision, half of the participants received three prompts, (a), "Recall the task and describe it briefly in your own words" (feed up), (b) "Explain how well your peer completed the task" (feed back), and (c) "Finally, make specific suggestions for how your peer can improve his or her case analysis" (feed forward). These prompts were accompanied by a sample solution illustrating how such feedback might look like (see Supplementary Infor-

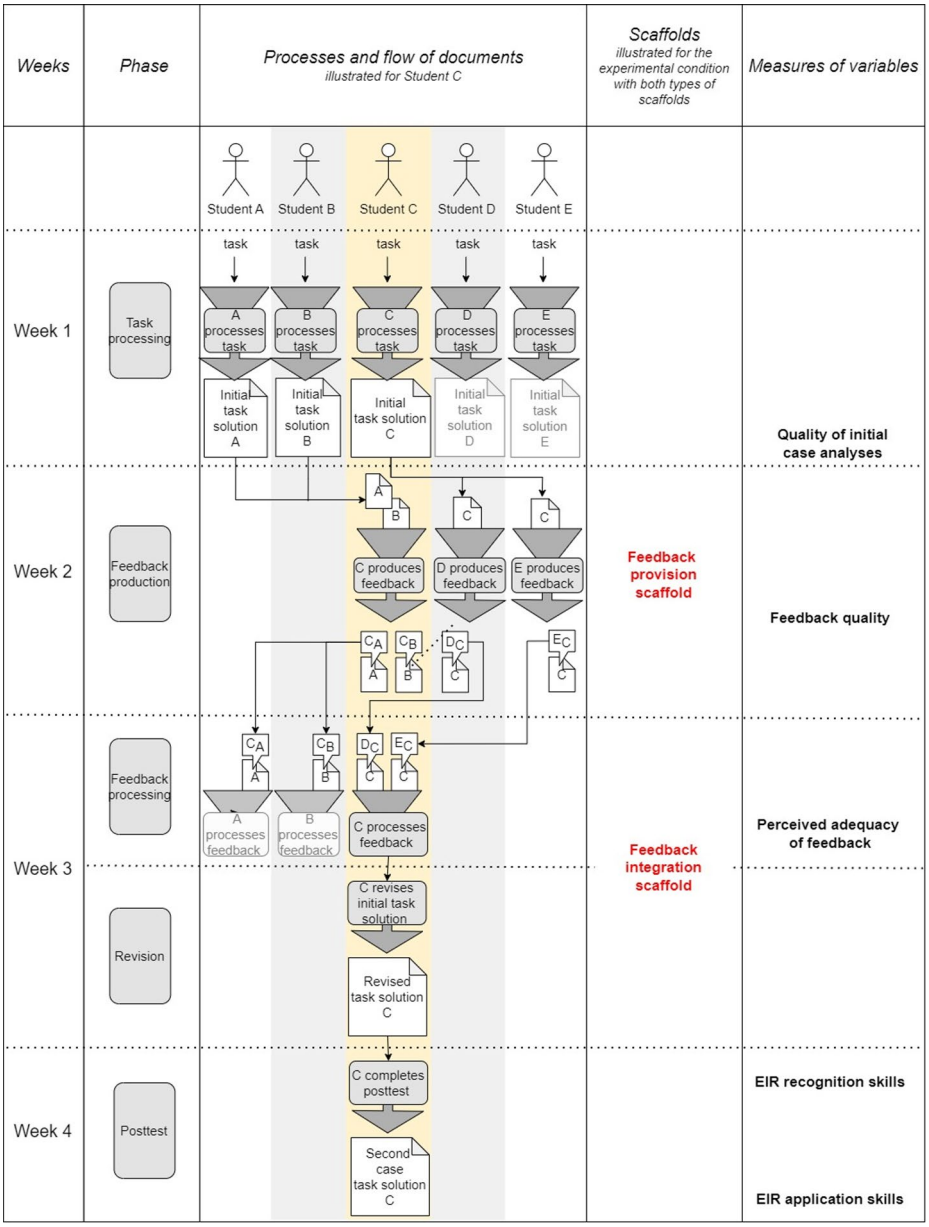

Fig. 1 Presentation of the study procedure within the experimental condition with both types of scaffolds. *Note.* Adapted from "Do students' beliefs and orientations toward peer feedback predict peer feedback quality and perceptions?" by M. Greisel, J., Hornstein, and I. Kollar, 2025, *Studies in Educational Evaluation, 84*, 101438. https://doi.org/10.1016/j.stueduc.2024.101438

mation, Appendix C). Participants who did not receive these scaffolds were asked to provide feedback without guidance, that is, they were only told to "please provide feedback on the first three problems identified by your peers". Later on, participants received feedback from peers who were from the same condition, that is, participants who were scaffolded in how to provide feedback also received feedback from peers who were scaffolded in the same way, while participants who were not scaffolded in providing feedback received feedback from peers who were also not scaffolded during feedback provision.

Regarding feedback reception and revision, participants received either prompts that explicitly guided them in integrating the feedback they received from two peers (e.g., "Mark statements that contradict each other with a third color (red)", see Table 1 for all prompts), or a single non-specific prompt ("Please revise your case analyses based on the feedback messages"). The prompts served as a step-by-step guide for integrating the peer-feedback messages (e.g., List & Alexander, 2019): First, participants were instructed to read both feedback messages. Second, they were asked to highlight instances in which the two feedback messages concurred in green, complementary instances in yellow, and instances in which they contradicted each other in red. Third, they were asked to revise their case analysis based on the two feedback messages. Fourth, they should highlight the changes in the same colors as in the second step and check whether they missed any criticism or suggestion. In a sample solution, participants were shown how the prompts could be implemented in practice when integrating the two peer-feedback messages (see Supplementary Information, Appendix D). Participants who were not in the feedback integration condition, in contrast, were asked to revise their initial case analyses without support.

Dependent Variables

Feedback quality. We coded feedback quality by aid of a category system with ten dimensions we derived from the literature (e.g., Hattie & Timperley, 2007; Langer et al., 2019; Nelson & Schunn, 2008; Patchan et al., 2016; Yu & Schunn, 2023). To end up with one single quality indicator, we used the total feedback message including all three problem analyses as the unit of analysis. This meant that the data was not segmented. This coding scheme was applied to analyze each feedback message for all problem analyses together and can be seen in the Supplementary Information (Appendix E).

First, each feedback message was coded with respect to eight structural criteria: In accordance with Hattie and Timperley (2007), we coded (1) the degree to which the task criteria were fulfilled (*feed up*); (2) whether feedback providers indicated when the solution

Table 1 Feedback integration prompts

1. Read: "First read both feedback messages one after the other."
2. Compare: "Mark similarities between the feedback messages with one color (green). Mark complementary statements that are only found in one of the feedback messages with a different color (yellow). Mark contradictory statements with a third color (red)."
3. Process: "Revise your own case analysis using the feedback messages."
4. Check: "In your own case analysis, mark the places where you have adapted your case analysis. Use the color you used to mark the corresponding statement in the feedback messages. Check whether all the marked criticisms or suggestions for improvement from the feedback messages are reflected in your case analysis and amend your work if necessary."

Note. These prompts appeared after participants in the feedback integration scaffold condition received the two peer-feedback messages on their initial case analysis

met expectations (*concordant feed back*); (3) when it fell short of the goal (*discrepant feed back*); and (4) whether it included recommendations for improvement (*feed forward*). Furthermore, we coded whether the feedback message (5) included details about the case analysis procedure (*process level*), and whether it (6) included further explanations or examples (*additional stimulation;* Langer et al., 2019) and (7) concepts from the associated theory materials in line with the conclusion from the literature that detailed feedback is preferable over simple verification (*reference to theory*; e.g., Nelson & Schunn, 2008). We further (8) used the number of characters in the feedback as an indicator for the likelihood for *feedback elaboration*, because we assume that the longer the feedback message is, the more likely it is that the feedback features it contains, will be spelled out in detail (cf. Patchan et al., 2016; Yu & Schunn, 2023). Second, we coded each feedback message with respect to its content quality, thereby using two categories: (9) *Relevance* refers to whether the feedback addressed all relevant points of the initial case analysis, and (10) *accuracy of content* refers to whether the feedback message was generally correct in relation to the initial case analysis.

Two independent raters (blind to condition) applied this coding scheme—after finalizing it through multiple rounds of independent coding and subsequent discussion—to 10% of the feedback messages, reaching satisfactory to excellent interrater agreement (Gwet's $AC1 = 0.66$ to 0.94). For the 10% we randomly selected one of the coders' coding. After that, the remaining 90% of the data were equally distributed between the two raters. Next, we ran an exploratory factor analysis with the eight structural criteria to see whether we could differentiate between different facets of structural feedback quality. This analysis did not suggest more than one primary factor (based on Kaiser-Guttman-Criterion, Scree Plot, and Map-Test). To arrive at a single structural quality score that weighs the different indicators in a non-arbitrary way, we used a one-factor confirmatory factor analysis to obtain the factor scores using the regression method and saved them as a variable (see Table 2 for factor loadings and standard errors). Content aspects (relevance and accuracy) were not included in the factor analysis, because these two categories are more dependent on subject-matter related knowledge than on general peer-feedback skills and therefore represent completely different constructs than the other categories. Therefore, in accordance with Nelson and Schunn

Table 2 Factor loadings and standard errors of the One-Factor confirmatory factor analysis to determine the structural feedback quality indicated by different structural criteria

Quality of Feedback 1	Factor loading	Standard error
Feed Up Aims	0.479	0.070
Feed Back Concordance	0.035	0.025
Feed Back Discrepancy	0.357	0.048
Feed Forward	0.459	0.043
Process	0.156	0.029
Additional Stimulation	0.168	0.050
Reference to Theory	0.362	0.050
Feedback Elaboration	0.687	0.081
Quality of Feedback 2		
Feed Up Aims	0.463	0.069
Feed Back Concordance	0.055	0.025
Feed Back Discrepancy	0.356	0.045
Feed Forward	0.387	0.044
Process	0.177	0.027
Additional Stimulation	0.253	0.045
Reference to Theory	0.475	0.041
Feedback Elaboration	0.766	0.069

(2008), we decided to give each of the two content criteria the same weight as the structural factor score. To arrive at an overall feedback quality score, we averaged the relevance, the accuracy, and the structural quality factor score.

EIR recognition skills. One week after participants had integrated and implemented the feedback they received from their peers, we asked them to complete a posttest. To measure participants' EIR recognition skills, we developed a test that presented three short case vignettes, each containing one instructional problem that could be solved by aid of cognitive load theory (Sweller, 1999) and/or the ICAP model (Chi & Wylie, 2014). Alongside each vignette, three different problem analyses were presented that differed in how well they utilized the five steps of evidence-informed reasoning. Participants had to sort the analyses from best to worst. In the best problem analysis, all five steps of evidence-informed reasoning were followed, whereas in the poorer problem analyses, steps were either merged or not followed at all. Participants received two points for a completely correct rank order, one point for an only slight deviation from the correct rank order, and zero points for a completely wrong rank order for each case vignette. We then calculated the mean score across the three vignettes and used it as the overall score for EIR recognition skills which can theoretically range from 0 to 2. Reliability was satisfactory (Cronbach's $\alpha = 0.64$). In addition, we found a significant positive correlation of the test scores with the quality of the initial case analysis ($r = .23$, $p < .001$) and with the quality of the revised case analysis ($r = .20$, $p < .01$), which can cautiously be interpreted as an indicator for the validity of our test. The correlations are very low, which is related to the fact that the two quality scores of the case analyses are used to map EIR skills, but we only refer to one sub facet (EIR recognition skills) in this test. This limits the maximum correlation that can be expected.

EIR application skills. To measure EIR application skills, we coded participants analyses of a second case vignette which they worked on in the posttest one week after they completed the peer-feedback process, again by aid of cognitive load theory (Sweller, 1999) and the ICAP model (Chi & Wylie, 2014), but without guidance. To assess these case analyses, we performed high-inference coding, using a rubric that included four levels from 0 = *poor quality* to 3 = *high quality* (for the coding scheme see Supplementary Information, Appendix F). The quality of the analysis of each of the three problems from the case vignette was coded separately, and we calculated an overall score from the three averages. Two independent raters, who were blind to condition, coded 10% of the data, after we finalized the coding scheme with multiple rounds of independent coding and discussions. They reached good interrater agreement (Gwet's $AC1 = 0.86$). After that, we randomly select one of the coders' coding, and the remaining 90% of the data were evenly distributed between the two coders.

Control Variables

With respect to RQ1, we included the quality of participants' own initial case analysis and the quality of the peers' initial case analysis as covariates. With respect to RQ2 and RQ3, we also used these variables as covariates, but also further variables related to the quality and adequacy of the provided and received feedback. The goal of including these variables as covariates was to reduce as much variance in these outcome variables as possible that would be caused by other factors than the experimental manipulation itself (in particular by processes that preceded the particular phase during which the dependent variable of interest was measured). That way, the remaining variance on the dependent variable can be

attributed to the experimental manipulation more clearly (for an exact description of which covariates were used for which research question, see "Statistical Analyses"). The covariates were measured as follows:

Quality of the initial case analyses. To assess the quality of the initial case analyses (included as covariate in the analyses on RQ1-3), we used the same coding procedure (Gwet's $AC1 = 0.99$) as for the second case (see above).

Feedback quality received and provided. These variables were described in the section "Dependent Variables" and were included as covariates in the analyses regarding RQ2 and RQ3.

Perceived adequacy of feedback. For each of the two feedback messages they received, participants answered nine items of the Feedback Perception Questionnaire (Strijbos et al., 2021) on a Likert scale from 1 = *completely disagree* to 9 = *completely agree*. These items capture three dimensions with three items each: fairness (e.g.: "I would consider this feedback fair."), usefulness (e.g.: "I would consider this feedback useful.") and acceptance (e.g.: "I would accept this feedback."). These items were aggregated to measure perceived adequacy (Peer 1: $\omega = 0.80$, Peer 2: $\omega = 0.82$). These covariates were included in the analyses regarding RQ2 and RQ3.

Statistical Analyses

All statistical analyses were performed in R [4.2.2]. We used the 'tidyverse' package [2.0.0] (Wickham et al., 2019) for data preparation and the 'psych' package [2.3.9] (Revelle, 2017) to calculate Cronbach alpha, McDonald's omega hierarchical, and an exploratory factor analysis. To answer our research questions, we calculated linear mixed models[2] with the scaffolds (regarding RQ1: feedback provision scaffold; regarding RQ2 and RQ3: feedback provision scaffold and feedback integration scaffold) as predictors, the dependent variable of interest as criterion variable, and the covariates described above. In these analyses, we utilized the REML-estimator using 'lme4' package [1.1–35.1] (Bates et al., 2015) and 'lmerTest' package [3.1-3] (Kuznetsova et al., 2017) to account for the two-level data structure (two feedback messages per participant). In addition, we used the 'lavaan' package [0.6–16] (Rosseel, 2012) to calculate the factor scores of the structural feedback quality (with WLSMV as estimator). We used the two existing scores based on the two feedback messages for the quality of the two case analyses a participant received to provide feedback on to create a mean score on which to base the analyses. We did the same for the quality of the feedback messages on the one hand and the adequacy of the feedback messages on the other.

Even though we had no hypotheses regarding the interaction effects (feedback provision scaffold*feedback integration scaffold), we included them as exploratory analyses, as significant interaction effects may affect how potential main effects should be interpreted. We set the alpha level at 5% for all calculations.

[2] Instead of ANCOVAs as described in the pre-registration, we used linear mixed models for the analyses because they are more flexible, provide more accurate results, and handle missing data better than ANCOVAs (cf. Brown, 2021).

Results

Preliminary Analyses

Table 3 shows the means, standard deviations, skewness, kurtosis, and correlations of the dependent and control variables. The quality of the feedback that participants provided correlated positively with EIR recognition skills and EIR application skills. In addition, as expected, the quality of one's own initial performance, which can be described as prior performance of EIR skills, correlates positively with EIR recognition skills and EIR application skills. Furthermore, the significant and positive correlation between EIR recognition skills and EIR application skills can cautiously be interpreted in a way that the two variables represent separate, but related components of the overarching construct "EIR skills".

Before computing linear mixed models to answer our research questions, we checked the assumptions for multiple regressions (linearity, normality, homoskedasticity, multicol-

Table 3 Means, standard deviations, skewness, kurtosis, and correlations

Variable	M	SD	Skewness	Kurtosis	1	2	3	4	5	6	7
Dependent Variables											
1. Feedback quality provided	0.33	0.32	−0.01	−0.25							
2. EIR recognition skills	0.65	0.47	0.33	−0.70	0.17**						
3. EIR application skills	1.12	0.88	0.42	−0.89	0.24**	0.21**					
Control Variables											
4. Coded quality of own initial case analyses	1.49	0.77	−0.03	−0.81	0.24**	0.24**	0.47**				
5. Coded quality of peers' initial case analyses	1.51	0.60	−0.14	−0.39	0.14*	0.04	0.13*	0.08			
6. Feedback quality received	0.36	0.34	0.21	0.40	0.10	−0.00	0.04	0.05	−0.06		
7. Perceived adequacy of feedback provided	7.29	1.25	−0.82	0.08	−0.00	0.03	−0.04	0.05	0.06	0.11	
8. Perceived adequacy of feedback received	7.44	1.23	−1.27	2.37	0.15*	−0.00	0.01	0.06	−0.00	0.02	0.17*

Note. M and *SD* are used to represent mean and standard deviation, respectively

* $p < .05$. ** $p < .01$

linearity, outliers) for all three research questions. Preliminary scatter plots showed that all relationships were linear. The residuals had a normal distribution and demonstrated homoscedasticity. Variance inflation factors for all predictors were close to one, the lowest feasible value, and far from larger than ten, which would suggest multicollinearity. We used Cook's distance, hat values, and studentized residuals to identify potential outliers, which we then displayed in influence plots. We then calculated the analyses both with and without the potential outliers. Since all potential outliers reflect valid values and their exclusion did not affect the significance of any of the effects we investigated, we decided not to exclude them from the analyses (see https://osf.io/vsbqd/?view_only=7657506b6a1244cb8e0d562e4e9ce8d0 https://osf.io/vsbqd/?view_only=3889d5bb4c274c0387be135a0b3306ff for the exact results of these analyses).

Research Question 1: Effects on Feedback Quality Provided

In this RQ, we refer only to the difference between the conditions with and without feedback provision scaffold because the feedback integration scaffold was not yet available at this time. Descriptively, for *the quality of the feedback that participants provided*, the condition with the feedback provision scaffold reached higher scores, than the condition without the feedback provision scaffold (see Table 4).

To test H1, we fitted a linear mixed model to predict the *quality of the feedback that participants provided* with the feedback provision scaffold (present vs. absent) as predictor, and the quality of participants' own and their peers' initial case analyses as covariates (see Table 5). The model explained a statistically significant and substantial proportion of variance. The effect of the feedback provision scaffold was statistically significant and positive. Thus, Hypothesis 1 was confirmed.

Research Question 2: Effects on EIR Recognition Skills

Descriptively, we found the highest mean score for EIR recognition skills in the feedback integration scaffold only condition (see Table 6), followed by the feedback provision scaffold only condition. Next were participants who received no scaffolds, followed by participants who received both scaffolds. However, these differences were only marginal, and all scores were rather low.

To test H2a and H2b, we fitted a linear mixed model to predict participants' *EIR recognition skills* as measured in the posttest with the feedback provision scaffold and the feedback integration scaffold as predictors. Feedback quality provided and received, perceived adequacy of the feedback received, quality of participants' own case analyses, and quality of their peers' case analyses were included as covariates. The model explained a statistically significant, but small proportion of variance (see Table 7). Neither the main effect of the feedback provision scaffold, nor of the feedback integration scaffold were significant. Thus, H2a and H2b were not supported by our results. Instead, the statistically significant negative

Table 4 Means and standard deviations for feedback quality per condition

	Feedback provision scaffold			
	Without ($n = 128$)		With ($n = 126$)	
	M	***SD***	***M***	***SD***
Feedback quality provided	0.24	0.29	0.44	0.31

Table 5 Results of the linear mixed model for the feedback quality provided

Variable	Estimates	CI	*p*	std. Beta	Std. CI
(Intercept)	−1.20	−1.61, −0.79	**<0.001**	−0.33	−0.50, −0.17
Predictor					
Feedback provision scaffold	0.68	0.45, 0.91	**<0.001**	0.78	0.45, 0.91
Control variables					
Coded quality of own initial case analysis	0.24	0.05, 0.44	**0.014**	0.15	0.03, 0.26
Coded quality of peers' initial case analyses	0.33	0.17, 0.48	**<0.001**	0.25	0.13, 0.36
Observations	239				
R^2/ R^2 adjusted	0.188/0.178				

Table 6 Means and standard deviations for EIR recognition skills per condition

	Feedback provision scaffold							
	Without				With			
	Feedback integration scaffold							
	Without (*n* = 66)		With (*n* = 62)		Without (*n* = 64)		With (*n* = 62)	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
EIR recognition skills	0.62	0.49	0.70	0.47	0.68	0.43	0.59	0.51

Table 7 Results of the linear mixed model for EIR recognition skills

Variable	Estimates	CI	*p*	std. Beta	Std. CI
(Intercept)	0.60	0.14, 1.05	**0.010**	−0.09	−0.35, 0.18
Predictor					
Feedback provision scaffold	0.12	−0.07, 0.32	0.206	0.26	−0.15, 0.68
Feedback integration scaffold	0.13	−0.04, 0.30	0.144	0.27	−0.09, 0.64
Interaction					
Feedback provision scaffold*feedback integration scaffold	−0.35	−0.61, −0.09	**0.009**	−0.75	−1.31, −0.19
Control variables					
Feedback quality provided	0.30	0.08, 0.52	**0.007**	0.21	0.06, 0.36
Feedback quality received	−0.06	−0.28, 0.17	0.622	−0.04	−0.20, 0.12
Perceived adequacy of the feedback received	−0.03	−0.08, 0.03	0.355	−0.07	−0.20, 0.07
Coded quality of own case analyses	0.11	0.02, 0.20	**0.014**	0.18	0.04, 0.32
Coded quality of peers' case analyses	−0.03	−0.14, 0.08	0.597	−0.04	−0.18, 0.10
Observations	208				
R^2/R^2 adjusted	0.101/0.065				

interaction effect indicates that the feedback provision scaffold only had a favorable impact when no additional integration scaffold was presented, and vice versa.

Research Question 3: Effects on EIR Application Skills

Descriptively, we observed that the participants from the condition with both scaffolds (see Table 8) reached the highest scores for EIR application skills. This is followed by the condition with feedback integration scaffold only and the condition without scaffolds. We found the lowest value for the condition with feedback provision scaffold only.

To test H3a and H3b, we fitted a linear mixed model to predict participants' *EIR application skills* as criterion variable with the feedback provision scaffold and the integration scaffold as predictors. As covariates, we included feedback quality provided and received, perceived adequacy of the feedback received, as well as the coded quality of own and peers' case analyses. The model explained a statistically significant and moderate proportion of variance (see Table 9). The effects of the feedback provision scaffold and the feedback integration scaffold were both statistically insignificant. Furthermore, the interaction effect was negative, but statistically not significant either. Thus, H3a and H3b were not confirmed.

Discussion

In this study, we applied peer-feedback to help pre-service teachers acquire EIR skills, which is an important goal in pre-service teacher education (e.g., Kollar et al., 2023). We therefore had pre-service teachers analyze authentic classroom cases using educational evidence, and asked them to provide feedback to their peers' analyses, receive and process feedback from their peers on their own analyses, and to use this feedback when revising their initial analyses. In this context, our goal was to investigate the effects of a feedback provision scaffold and a feedback integration scaffold on pre-service teachers' (a) feedback quality, (b) EIR recognition skills, and (c) EIR application skills.

Research Question 1: Effects on Feedback Quality Provided

With respect to feedback quality, the results indicated that, as expected, the feedback provision scaffold had large, significant, and positive effects on feedback quality. Thus, it seems that feedback provision scaffolds as the one we designed are powerful tools that enable pre-service teachers to formulate feedback messages that have a strong potential to support their peers' learning processes. These findings are consistent with previous literature, indicating that structuring feedback provision with prompts and related sample solutions that help learners formulate feedback that includes information on feed up, feed back, and feed

Table 8 Means and standard deviations for EIR application skills per condition

	Feedback provision scaffold							
	Without				With			
	Feedback integration scaffold							
	Without ($n=66$)		With ($n=62$)		Without ($n=64$)		With ($n=62$)	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
EIR application skills	1.04	0.80	1.15	0.91	1.01	0.88	1.31	0.90

Table 9 Results of the linear mixed model for the EIR application skills

Variable	Estimates	CI	*p*	std. Beta	Std. CI
(Intercept)	0.26	−0.48, 1.00	0.492	−0.08	−0.33, 0.16
Predictor					
Feedback provision scaffold	0.04	−0.29, 0.36	0.826	0.04	−0.33, 0.42
Feedback integration scaffold	0.14	−0.16, 0.44	0.352	0.16	−0.18, 0.50
Interaction					
Feedback provision scaffold*feedback integration scaffold	−0.05	−0.49, 0.40	0.832	−0.06	−0.57, 0.46
Control variables					
Feedback quality provided	0.37	−0.00, 0.75	0.053	0.14	−0.00, 0.28
Feedback quality received	−0.02	−0.41, 0.37	0.917	−0.01	−0.15, 0.14
Perceived adequacy of the feedback received	−0.01	−0.11, 0.08	0.785	−0.02	−0.15, 0.11
Coded quality of own case analyses	0.33	0.10, 0.56	**0.006**	0.28	0.08, 0.47
Coded quality of peers' case analyses	0.20	−0.04, 0.44	0.096	0.17	−0.03, 0.36
Observations	206				
R^2/R^2 adjusted	0.231/0.200				

forward can have positive effects on feedback quality (e.g., Gan & Hattie, 2014; Gyamfi et al., 2024; Latifi et al., 2021). That it is possible to increase the quality of feedback messages by aid of scaffolding is of great importance for the further peer-feedback process, as high-quality feedback provides more useful information to feedback receivers to help them revise their initial solutions and that way expand on their knowledge (e.g., Hattie & Timperley, 2007). However, providing high-quality feedback may also be helpful for the feedback providers themselves, as they need to deeply process their peers' initial solutions and relate them to educational evidence (e.g., Carless & Boud, 2018; Cho & Cho, 2011). Also, the scaffolding we used is rather easy to implement in practice, especially when comparing it to training interventions that precede the actual peer-feedback process (see Man et al., 2024). Our results indicate that well-designed scaffolds that are integrated into the learning environment at the right place can also have large positive effects. Further, the prompts we used strongly focused on the structural rather than content-related aspects of feedback. This is another advantage because they can easily be used in other domains and contexts without having to make major adjustments (e.g., Hornstein et al., 2025).

Research Question 2: Effects on EIR Recognition Skills

RQ2 focused on the effects of the two scaffolds on pre-service teachers' EIR recognition skills, which—based on skill acquisition theories (e.g., Anderson & Lebiere, 1998) and research on professional vision (e.g., Gegenfurtner et al., 2020)—can be regarded as a first, early indicator of EIR skills. Our results showed neither a significant main effect of the feedback provision scaffold nor a significant main effect of the feedback integration scaffold. In contrast, we observed a significant negative interaction effect, which indicates that the two scaffolds worked better alone than when they were combined.

At least two explanations for this result are possible: First, from a utilization deficiency perspective (e.g., Bjorklund et al., 1997), participants might not have internalized the strategies well enough to handle peer-feedback processes effectively yet. In other words, applying the new strategies that the two scaffolds prompted may take up too many cognitive resources to convert their potential usefulness into beneficial effects when it comes to recognizing evidence-informed reasoning. This might in particular be problematic if the task is too complex for learners, especially when they have insufficient prior domain knowledge (e.g., Alqassab et al., 2018; van Zundert, Sluijsmans et al., 2012), which might have been the case with our participants. This means, that learning from peer-feedback might take time and practice until it really pays off, which, in turn, would be an argument in favor of an implementation of extensive training interventions, as mentioned above (e.g., Li et al., 2020; van Zundert et al., 2010). However, training interventions are often time-consuming, both with respect to their development and their implementation in practice (cf. Man et al., 2024).

A second explanation for the negative interaction effect could be that the two scaffolds might have triggered cognitive processes that might have interfered with each other (see Tabak, 2004): While the feedback provision scaffold led participants to focus on the structural components of feedback, the integration scaffold directed their attention to content-specific similarities and dissimilarities between peer-feedback messages, regardless of the structure of the feedback. As a result, the feedback integration scaffold might have diverted participants' attention away from the required analysis method. Future research might therefore investigate if interference processes can be avoided if feedback integration scaffolds are designed in a way that they also focus on structural (instead of content-related) feedback elements. This is also related to the orientation of the two scaffolds: If we look at them more closely, we see that they do not explicitly refer to evidence-informed reasoning, but rather to how to provide structurally good feedback and how to integrate multiple peer-feedback messages received. This focus of the scaffolds may have resulted in insufficient support for EIR recognition skills.

At a more general level, the negative interaction effect of the two scaffolds reflects the challenges that are associated with an optimal coordination of multiple scaffolds (see Quintana et al., 2004). Future research should thus continue to find ways how to optimally combine different scaffolds to yield the best possible effects of peer-feedback.

Interestingly, we found a significant effect of the covariate feedback quality provided on EIR recognition skills. This could mean that providing better feedback leads to stronger EIR processing, which may be helpful for EIR recognition (e.g., van Popta et al., 2017). However, it should be noted that this effect is correlative and not causal, so it could also be a spurious correlation. That is, more proficient participants write better feedback and at the

same time recognize better evidence-informed reasoning case analyses. This correlation should be further explored in future research.

Research Question 3: Effects on EIR Application Skills

Research Question 3, finally, addressed the effects of the feedback provision scaffold and the feedback integration scaffold on pre-service teachers' EIR application skills. Contrary to our expectations, we found no such effects, neither for the feedback provision scaffold nor for the feedback integration scaffold, nor for their interaction. Thus, it seems to be difficult to produce effects on EIR application skills by aid of the scaffolding we used. A straightforward interpretation of this finding is that skill acquisition—as suggested by related theories such as ACT-R (Anderson & Leibiere, 1998)—usually first requires the development of a declarative representation of the to-be-learnt skill, before this declarative knowledge can be proceduralized through practice. Applied to our study, since we found no significant effects of the scaffolds on EIR recognition skills, it is not surprising that the scaffolds also do not have a significant effect on EIR application skills. However, to actually perform evidence-informed reasoning on a new case vignette might require more proceduralized knowledge, and: more opportunities for practice. Along these lines, as expected, Table 3 shows a significant correlation between the initial quality of participants' own case analyses (i.e., participants prior performance of EIR skills) and their EIR application skills. This means that participants who already had a high level of performance in EIR application skills from the outset also show better performance afterwards when analyzing the second case. For the design of scaffolds for peer-feedback in the context of pre-service teachers' evidence-informed reasoning, this may indicate that it is important to first help pre-service teachers build up a declarative representation of EIR skills, and only then have them engage in evidence-informed analyses of case vignettes by aid of peer-feedback (e.g., van Zundert, Könings et al., 2012). If pre-service teachers are asked to apply the newly learned skills, this may take up too many cognitive resources and thus produce the aforementioned utilization deficit (e.g., Bjorklund et al., 1997). It might well be that even though our scaffolds were intended to reduce task complexity, this reduction was not enough to help participants use EIR application skills at a high level (e.g., van Zundert, Sluijsmans et al., 2012). Along these lines, and supported by findings by van Zundert, Könings et al. (2012), it might be more helpful to first provide opportunities to acquire domain-specific knowledge (in our case: knowledge about educational theories and related empirical evidence), before having them engage in peer-feedback processes in which this knowledge is needed.

We also should take into account that the feedback scaffolds we used did not directly address the correctness of the comments and suggestions provided by the peers. Instead, they focused more on the structural components of feedback (see e.g., Hattie & Timperley, 2007), and the integration of multiple peer-feedback messages. Thus, it might well be that some participants incorporated incorrect feedback into their revisions, which might have harmed their acquisition of EIR application skills. Future studies might therefore develop and evaluate the effects of more content-related peer-feedback scaffolds (e.g., scaffolds that provide guidance on how to distinguish correct from incorrect feedback).

Limitations

Our study is certainly not without limitations. First, we do not know how participants dealt with the two peer-feedback messages and which parts of the feedback they used for the revision of their initial case analyses. In other words, it is unclear how exactly the integration went about. Yet, feedback integration can be very complex if learners receive more than one peer-feedback message and these peer-feedback messages contradict each other (cf., Wu & Schunn, 2021). It would certainly be interesting to investigate the effects of the feedback integration scaffold further, possibly using qualitative methods.

Second, as our scaffolds combined different instructional aides to maximize learning benefits based on best practices and empirical evidence from research on prompting and sample solutions (e.g., Gyamfi et al., 2024), we cannot tax the isolated contribution of the prompts and the sample solutions on the dependent variables we investigated. Future studies might try to disentangle these effects.

Third, our study focused on *instructional methods* that support evidence-informed reasoning by pre-service teachers. However, also other, and in particular learner-related factors, such as pre-service teachers' motivation, may play a crucial role. As described above, these factors might also explain why the combination of feedback provision scaffold and feedback integration scaffold did not work ideally. Thus, even though research has already been conducted that investigates the influence of individual learner characteristics during peer-feedback (e.g., Aben et al., 2019; Greisel et al., 2025), future research might further expand on this, especially with respect to uncovering possible interactions between learner prerequisites and instructional variables.

Conclusions

Without downplaying these limitations, our study demonstrates that enhancing feedback provision with a scaffold that helps pre-service teachers formulate high-quality feedback can be an effective way to improve the quality of peer-feedback messages (e.g., Gyamfi et al., 2024; Hornstein et al., 2025). We also found a negative interaction effect of the feedback provision scaffold and the feedback integration scaffold with respect to EIR recognition skills. This is a surprising and interesting finding, as our study is among the first to even tackle the question how learners can be instructionally supported during feedback provision as well as during feedback integration. Though, more research is needed to identify powerful ways to support learners during feedback provision as well as feedback integration.

In addition, we saw that to produce effects also on pre-service teachers' EIR application skills, further and/or alternative scaffolds seem to be necessary, possibly especially ones that help pre-service teachers build up prior domain-specific knowledge before actually engaging in peer-feedback (e.g., van Zundert, Könings et al., 2012). Also, further learner-related factors (e.g., understanding of received peer-feedback messages, agreement with received peer-feedback messages, perceptions about peer-feedback) may influence whether or not pre-service teachers benefit from peer-feedback on their EIR skills (e.g., Aben et al., 2019; Strijbos & Müller, 2014; Wu & Schunn, 2020), and more research on this topic is needed. The same is true regarding how to best combine feedback provision scaffolds with feedback integration scaffolds. Here, our results indicate that it is important that the targeted processes

do not interfere with each other. It might thus be promising to investigate how feedback provision scaffolds on the one hand and feedback integration scaffolds on the other should best be synchronized to achieve synergistic scaffolding effects (Tabak, 2004). Additionally, the context in which peer-feedback occurs plays a crucial role in determining its effectiveness: Learners need enough prior knowledge to be able to provide correct feedback and to be able to decide whether the feedback they receive is correct in terms of content (e.g., Alqassab et al., 2018). Moreover, the complexity of the task should be appropriate for the use of peer-feedback to facilitate meaningful engagement (e.g., van Zundert, Könings et al., 2012). Ensuring both optimal scaffold synchronization and a suitable peer-feedback context may therefore be key to enhancing learning outcomes.

Even though these questions have to remain open, our study provides important evidence for the usefulness and applicability and design of peer-feedback in pre-service teacher education when the goal is to help pre-service teachers become more competent in evidence-informed reasoning.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11251-025-09745-4.

Author contributions Conceptualization: Julia Hornstein, Martin Greisel, Ingo Kollar; Methodology: Julia Hornstein, Martin Greisel, Ingo Kollar; Formal analysis and investigation: Julia Hornstein, Martin Greisel; Writing - original draft preparation: Julia Hornstein; Writing - review and editing: Martin Greisel, Ingo Kollar; Funding acquisition: Ingo Kollar; Supervision: Ingo Kollar.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Stiftung Innovation in der Hochschullehre (project "Facilitating Competence Development through Authentic, Digital, and Feedback-Based Teaching-Learning Scenarios", grant number FBM2020).

Declarations

Financial interests The authors declare they have no financial interests.

Non-financial interests Ingo Kollar is member of the editorial advisory board for Instructional Science.

References

Aben, J. E. J., Dingyloudi, F., Timmermans, A. C., & Strijbos, J.-W. (2019). Embracing Errors for Learning: Intrapersonal and Interpersonal Factors in Feedback Provision and Processing in Dyadic Interactions. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The impact of feedback in higher education: Improving assessment outcomes for learners* (pp. 107–125). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-25112-3_7

Alemdag, E., & Yildirim, Z. (2022). Effectiveness of online regulation scaffolds on peer feedback provision and uptake: A mixed methods study. *Computers & Education*, *188*, 104574. https://doi.org/10.1016/j.compedu.2022.104574

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, *33*(1), 11–30. https://doi.org/10.1007/s10212-017-0342-0

Anderson, J. R., & Lebiere, C. J. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315805696

Baker, K. M. (2016). Peer review as a strategy for improving students' writing process. *Active Learning in Higher Education*, *17*(3), 179–192. https://doi.org/10.1177/1469787416654794

Barzilai, S., Zohar, A. R., & Mor-Hagani, S. (2018). Promoting integration of multiple texts: A review of instructional approaches and practices. *Educational Psychology Review*, *30*(3), 973–999. https://doi.org/10.1007/s10648-018-9436-8

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software,67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Berndt, M., Strijbos, J. W., & Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *European Journal of Psychology of Education*, *33*(1), 31–49. https://doi.org/10.1007/s10212-017-0343-z

Bjorklund, D. F., Miller, P. H., Coyle, T. R., & Slawinski, J. L. (1997). Instructing children to use memory strategies: Evidence of utilization deficiencies in memory training studies. *Developmental Review*, *17*(4), 411–441. https://doi.org/10.1006/drev.1997.0440

Boscolo, P., Arfé, B., & Quarisa, M. (2007). Improving the quality of students' academic writing: An intervention study. *Studies in Higher Education,32*(4), 419–438. https://doi.org/10.1080/03075070701476092

Bråten, I., Anmarkrud, Ø., Brandmo, C., & Strømsø, H. I. (2014). Developing and testing a model of direct and indirect relationships between individual differences, processing, and multiple-text comprehension. *Learning and Instruction*, *30*, 9–24. https://doi.org/10.1016/j.learninstruc.2013.11.002

Brown, V. A. (2021). An introduction to linear Mixed-Effects modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592096035. https://doi.org/10.1177/2515245920960351

Bürgermeister, A., Glogger-Frey, I., & Saalbach, H. (2021). Supporting peer feedback on learning strategies: Effects on self-efficacy and feedback quality. *Psychology Learning & Teaching,20*(3), 383–404. https://doi.org/10.1177/14757257211016604

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology*, *100*(1), 209–222. https://doi.org/10.1037/0022-0663.100.1.209

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

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

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

Csanadi, A., Kollar, I., & Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *European Journal of Psychology of Education*, *36*(1), 147–168. https://doi.org/10.1007/s10212-020-00467-4

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. https://doi.org/10.1007/s10648-019-09510-3

Ferguson, L. E., & Bråten, I. (2022). Unpacking pre-service teachers' beliefs and reasoning about student ability, sources of teaching knowledge, and teacher-efficacy: A scenario-based approach. *Frontiers in Education*, *7*, 975105. https://doi.org/10.3389/feduc.2022.975105

Gan, M. J., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science,42*(6), 861–878. https://doi.org/10.1007/s11251-014-9319-4

Gegenfurtner, A., Lewalter, D., Lehtinen, E., Schmidt, M., & Gruber, H. (2020). Teacher expertise and professional vision: Examining knowledge-based reasoning of pre-service teachers, in-service teachers, and school principals. *Frontiers in Education,5*, 59. https://doi.org/10.3389/feduc.2020.00059

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, *52*, 315–325. https://doi.org/10.1016/j.chb.2015.06.019

Gorard, S. (Ed.). (2020). *Getting Evidence into Education: Evaluating the Routes to Policy and Practice* (1st ed.). Routledge. https://doi.org/10.4324/9780429290343

Greisel, M., Hornstein, J., & Kollar, I. (2025). Do students' beliefs and orientations toward peer feedback predict peer feedback quality and perceptions? *Studies in Educational Evaluation, 84*. https://doi.org/10.1016/j.stueduc.2024.101438

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

Gyamfi, G., Hanna, B. E., & Khosravi, H. (2024). Impact of an instructional guide and examples on the quality of feedback: Insights from a randomised controlled study. *Educational Technology Research and Development*, *72*(3), 1419–1437. https://doi.org/10.1007/s11423-024-10346-0

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

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

Helleve, I., Eide, L., & Ulvik, M. (2023). Case-based teacher education preparing for diagnostic judgement. *European Journal of Teacher Education,46*(1), 50–66. https://doi.org/10.1080/02619768.2021.1900112

Hornstein, J., Keller, M. V., Greisel, M., Dresel, M., & Kollar, I. (2025). Enhancing the Peer-Feedback Process Through Instructional Support: A Meta-Analysis. *Educational Psychology Review, 37*(2). https://doi.org/10.1007/s10648-025-10017-3

Hsiao, Y. P., & Rajagopal, K. (2023). Support Student Integration of Multiple Peer Feedback on Research Writing in Thesis Circles. In: O. Noroozi, & B. De Wever (Eds.*), The Power of Peer Learning. Social Interaction in Learning and Development* (pp. 47–70). Springer, Cham. https://doi.org/10.1007/978-3-031-29411-2_3

Huisman, B., Saab, N., Van Den Broek, P., & Van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education,44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Ion, G., Sánchez Martí, A., & Agud Morell, I. (2018). Giving or receiving feedback: Which is more beneficial to students' learning? *Assessment & Evaluation in Higher Education*, *44*(1), 124–138. https://doi.org/10.1080/02602938.2018.1484881

Jurkowski, S. (2018). Do question prompts support students in working with peer feedback? *International Journal of Educational Research*, *92*, 1–9. https://doi.org/10.1016/j.ijer.2018.07.003

Kollar, I., Greisel, M., Krause-Wichmann, T., & Stark, R. (2023). Editorial: Evidence-informed reasoning of pre- and in-service teachers. *Frontiers in Education*, *8*., Article 1188022. https://doi.org/10.3389/feduc.2023.1188022

König, J., Bremerich-Vos, A., Buchholtz, C., & Glutsch, N. (2020). General pedagogical knowledge, pedagogical adaptivity in written lesson plans, and instructional practice among preservice teachers. *Journal of Curriculum Studies*, *52*(6), 800–822. https://doi.org/10.1080/00220272.2020.1752804

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Langer, I., von Thun, S., F., & Tausch, R. (2019). *Sich verständlich ausdrücken* (11. Auflage) [*Express yourself clearly* (11th Edition)]. Ernst Reinhardt Verlag. https://elibrary.utb.de/doi/book/10.2378/9783497611331

Latifi, S., Noroozi, O., & Talaee, E. (2021). Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes. *British Journal of Educational Technology*, *52*(2), 768–784. https://doi.org/10.1111/bjet.13054

Li, L., & Grion, V. (2019). The power of giving feedback and receiving feedback in peer assessment. *All Ireland Journal of Higher Education*, *11*(2). https://doi.org/10.62707/aishej.v11i2.413

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education,45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

List, A., & Alexander, P. A. (2019). Toward an Integrated Framework of Multiple Text Use. *Educational Psychologist, 54*(1), 20–39. https://doi.org/10.1080/00461520.2018.1505514

Lui, A. M., & Andrade, H. L. (2022). The next black box of formative assessment: A model of the internal mechanisms of feedback processing. *Frontiers in Education*, *7*, 751548. https://doi.org/10.3389/feduc.2022.751548

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing,18*(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Maier, J., & Richter, T. (2014). Fostering multiple text comprehension: How metacognitive strategies and motivation moderate the text-belief consistency effect. *Metacognition and Learning*, *9*(1), 51–74. https://doi.org/10.1007/s11409-013-9111-x

Man, D., Chau, M. H., Zainal, A. Z., & Rummy, J. T. (2024). Exploring the potential of brief peer review training in enhancing student feedback literacy. *International Review of Applied Linguistics in Language Teaching Advance Online Publication*. https://doi.org/10.1515/iral-2023-0278

Menz, C., Spinath, B., & Seifried, E. (2021). Where do pre-service teachers' educational psychological misconceptions come from? The roles of anecdotal versus scientific evidence. *Zeitschrift für Pädagogische Psychologie,35*(2–3), 143–156. https://doi.org/10.1024/1010-0652/a000299

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Routledge. https://doi.org/10.4324/9780203880869

Nelson, M. M., & Schunn, C. D. (2008). 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., & 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 and Evaluation in Higher Education,39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

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

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

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal Of The Learning Sciences,13*(3), 337–386. https://doi.org/10.1207/s15327809jls1303_4

Revelle, W. R. (2017). *psych: Procedures for Personality and Psychological Research.*

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

Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). IAP Information Age Publishing.

Rousseau, D. M., & Gunia, B. C. (2016). Evidence-based practice: The psychology of EBP implementation. *Annual Review of Psychology,67*, 667–692. https://doi.org/10.1146/annurev-psych-122414-033336

Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal*, *51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Shavelson, R. J. (2020). Research on teaching and the education of teachers: Brokering the gap. *Beiträge zur Lehrerinnen- und Lehrerbildung,38*(1), 37–53. https://doi.org/10.36950/bzl.38.2020.9310

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*(1), 21–31. https://doi.org/10.1080/00461520.2019.1611432

Sluijsmans, D. M. A., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education*, *27*(5), 443–454. https://doi.org/10.1080/0260293022000009311

Stadtler, M., Scharrer, L., Skodzik, T., & Bromme, R. (2014). Comprehending multiple documents on scientific controversies: Effects of reading goals and signaling rhetorical relationships. *Discourse Processes*, *51*(1–2), 93–116. https://doi.org/10.1080/0163853X.2013.855535

Strijbos, J. W., & Müller, A. (2014). Personale Faktoren Im Feedback Prozess [Personal factors in the feedback process]. In H. Ditton, & A. Müller (Eds.), *Feedback und rückmeldungen: Theoretische grundlagen, empirische befunde, Praktische Anwendungsfelder [Feedback and evaluation: theoretical foundations, empirical findings, practical implementation]* (pp. 83–134). Waxmann.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction,20*(4), 291–303. https://doi.org/10.1016/j.learninstruc.2009.08.008

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

Sweller, J. (1999). *Instructional design in technical areas*. ACER.

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *The Journal of the Learning Sciences*, *13*(3), 305–335. http://www.jstor.org/stable/1466940

van der Pol, J., van den Berg, B. A. M., Admiraal, W. F., & Simons, R. J. P. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education*, *51*(4), 1804–1817. https://doi.org/10.1016/j.compedu.2008.06.001

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

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

van Zundert, M. J., Könings, K. D., Sluijsmans, D., & van Merriënboer, J. (2012). Teaching domain-specific skills before peer assessment skills is superior to teaching them simultaneously. *Educational Studies*, *38*(5), 541–557. https://doi.org/10.1080/03055698.2012.654920

van Zundert, M. J., Sluijsmans, D. M. A., Könings, K. D., & van Merriënboer, J. J. (2012). The differential effects of task complexity on domain-specific and peer assessment skills. *Educational Psychology*, *32*(1), 127–145. https://doi.org/10.1080/01443410.2011.626122

Voss, T. (2022). Not useful to inform teaching practice? Student teachers hold skeptical beliefs about evidence from education science. *Frontiers in Education*, *7*, 976791. https://doi.org/10.3389/feduc.2022.976791

Walker, M. (2015). The quality of written peer feedback on undergraduates' draft answers to an assignment, and the use made of the feedback. *Assessment and Evaluation in Higher Education,40*(2), 232–247. https://doi.org/10.1080/02602938.2014.898737

Wecker, C., & Fischer, F. (2014). Where is the evidence? A meta-analysis on the role of argumentation for the acquisition of domain-specific knowledge in computer-supported collaborative learning. *Computers & Education,75*, 218–228. https://doi.org/10.1016/j.compedu.2014.02.016

Wichmann, A., Funk, A., & Rummel, N. (2018). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *European Journal of Psychology of Education*, *33*(1), 165–184. https://doi.org/10.1007/s10212-017-0348-7

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilkes, T., & Stark, R. (2023). Probleme evidenzorientierter Unterrichtspraxis: Anregungen und Lösungsvorschläge [Problems with evidence-based teaching practice. Suggestions and proposed solutions]. *Unterrichtswissenschaft, 51*, 289–313. https://doi.org/10.1007/s42010-022-00150-1

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, *60*, 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science*, *49*(3), 365–394. https://doi.org/10.1007/s11251-021-09546-5

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