Intern. J. Comput.-Support. Collab. Learn
https://doi.org/10.1007/s11412-023-09401-4

Does matching peers at finer-grained levels of prior performance enhance gains in task performance from peer review?

Zheng Zong[1] · **Christian D. Schunn**[2]

Received: 22 November 2022 / Accepted: 29 May 2023

Abstract
Online peer feedback has proven to be practically useful for instructors and to be useful for learning, especially for the feedback provider. Because students can vary widely in skill level, some research has explored matching reviewer and author by performance level. However, past research on the impacts of reviewer matching has found little effect but used a simple binary high–low approach, which may mask the relative benefits of performance matching. In the current study, we leveraged a large dataset involving three large biology courses implementing multiple assignments with online peer feedback. This large dataset enabled dividing students into four levels of relative task performance to tease apart the relative contributions of providing and receiving feedback within the 16 different author–reviewer performance pairings. The results reveal that changes in task performance over assignments attributable to reviewing experiences vary by these finer-grained prior performance distinctions. In particular, providing to students at the same performance level appears to be beneficial, and receiving feedback from students at the same level is helpful except for very low-performing students. A simulation was used to examine the combined effects of receiving and providing under different algorithms for assigning reviewers to documents. The simulations suggest a matching algorithm will produce overall better outcomes than a random assignment algorithm for students at each of the four performance levels.

Keywords Peer review · Peer feedback · Task performance · Achievement grouping · Simulation

✉ Zheng Zong
zongzheng@cupl.edu.cn

Christian D. Schunn
schunn@pitt.edu

[1] School of Law and Economics, China University of Political Science and Law, No. 25 Xitucheng Road., Haidian District, Beijing 100088, People's Republic of China

[2] Learning Research & Development Center, University of Pittsburgh, Pittsburgh, PA, USA

Published online: 16 August 2023

Introduction

Peer review involves students responding to work completed in the same class (Chien et al., 2020; Vanhorn et al., 2019), whether the student work be essays (Oppermann et al., 2019; Yim & Warschauer, 2017), video presentations (Min, 2016), design projects (Kim & Ketenci, 2019), or computer code (Seifert & Feliks, 2019; Wang & Sun, 2018). Peer review can involve evaluative ratings (often called peer assessment) as well as written or verbal comments provided to the author of the work (often called peer feedback) in-person or through computer-mediated forms. There are a wide number of reasons for wanting to include peer review in coursework. Pragmatically, particularly when implemented with supporting technologies, it allows for more assignments involving complex work because of unsustainable grading loads. More importantly, there are substantial learning benefits from participating in peer review as established in multiple meta-analyses (Chang et al., 2021; Huisman et al., 2019; Li et al., 2020), especially from the aspect of students providing feedback to peers (Zong et al., 2021).

An obvious but nontrivial issue in peer review is that there can be large differences within a course in the foundational skills of students that can influence the quality of the peer reviews (e.g., review length, rating accuracy, helpfulness for revision). Students often voice concerns that their peers are not competent enough to provide useful feedback (Shang, 2022), and that grading should be the instructor's task (Adachi et al., 2018; Dawson et al., 2019). Indeed, studies of peer review have revealed that the quality of peer feedback received by students sometimes depends upon the prior achievement of peers assigned to the reviewing task (e.g., Patchan & Schunn, 2015; Huisman et al., 2018), and the issues are about ignoring challenging problem topics or not offering solutions rather than offering incorrect advice (Patchan et al., 2018; Ramachandran et al., 2017; Wu & Schunn, 2021a).

However, although there are widespread concerns about differences in students' achievements influencing review quality, few studies have systematically examined the extent to which involving reviewers of higher or lower achievements enhances learning (i.e., whether differences in the contents of reviews result in observable changes in later task performance). Furthermore, since the difficulty of assignments will tend to vary greatly between more introductory versus more advanced courses, the concern is not for the absolute level of student task performance within a discipline, but rather the level of performance relative to others in the class: do students have sufficient understanding of the current task and the kinds of problems that students in that class have so that they are able to provide feedback that is useful; are peers discussing issues that are an appropriate focus for a given learner? Therefore, it is necessary to consider the relationship between students' own achievements and the achievements of their peers.

One potentially relevant classroom teaching strategy that might address variation in students' achievements is to use a matching strategy: group students according to achievement. In online peer feedback, it is relatively easy to create algorithms that group students by prior performance (Abrache et al., 2021). Outside of peer feedback, this matching approach has received some support in improving learning. For example, grouping students can improve students' confidence and concentration (Casserly et al., 2019) or team performance (Hastings et al., 2022; Wen et al., 2017). Note that we are not discussing grouping by general cognitive ability, such as whether students classified as gifted/talented or having special learning needs should be working in mixed or separated classrooms or groups.

Student grouping by task performance has been explored in peer feedback research, considering benefits of prior performance matching (e.g., Patchan et al., 2018; Huisman

et al., 2018) and in creating systems that implement matching (e.g., Abrache et al., 2021). However, the empirical research on grouping in peer feedback by prior performance has tended to focus on relatively gross student performance level distinctions (i.e., only a high/low binary distinction). Further, much of the research examined process characteristics (e.g., how much feedback was received, what kind of feedback was received) rather than the benefits of such grouping on changes in students' later task performance. The current work explores matching by four levels of performance rather than only two levels (using mean peer ratings as the measure of task performance), and it examines the impacts of various forms of match/mismatch (within both receiving and providing) on students' gradual improvements in later task performance across assignments. Understanding the benefits of mismatches as well as exact matching can support the development of new, more effective reviewer assignment algorithms. Further, understanding the separable benefits/challenges from both a providing and receiving perspective can illustrate areas in need of further computer supports during the peer feedback process.

Theoretical background

Zone of proximal development and peer review

In prior research, many theoretical approaches have been applied to investigating peer feedback, including theories of deliberate practice (e.g., Kellogg & Whiteford, 2009) and self-regulated learning (e.g., Alemdag & Yildirim, 2022; Yang et al., 2006). However, the most commonly cited theoretical foundation to educational applications of peer feedback is the Zone of Proximal Development, which is defined as "the distance between the actual developmental level as determined by independent problem-solving and the level of potential development as determined through problem-solving under adult guidance or in collaboration with a more capable peer" (Vygotsky, 1978, P. 86). In contrast to instructor feedback, the feedback from peers can be thought of as providing students with learning opportunities within their zone of proximal development (De Guerrero & Villamil, 2000; Warwick & Maloch, 2003; Yu & Hu, 2017). That is, peer feedback activities ask students to evaluate and provide constructive criticism on topics/issues that are often just beyond what they can reliably do on their own (Wu & Schunn, 2020). However, this simple conceptualization treats all students within a class as functionally identical: they are all working with a shared zone of proximal development. In many classroom contexts, there can be large differences in the achievement levels of peers, meaning that particular pairings of students within a class may not be within the right zone. Students and instructors often worry whether students with lower current performance can provide meaningful evaluations and feedback to students with higher current performance (Brown et al., 2016; Hsiao & Brusilovsky, 2011).

Beyond peer feedback, many studies have examined the quality of interaction within collaborators of different prior achievements, including interactions between experts (very high-level achievement) and novices (very low-level achievement), as well within interactions between novices and novices. Some studies found that both "expert" companions and "novice" companions can support student learning by asking questions, suggesting possible solutions, and encouraging practice through repetition (Shooshtari & Mir, 2014; Storch, 2018). However, some studies have also found that "expert" companions are indispensable. For example, in the absence of "expert" companions, learners' learning of content may be poor (Schmid & Finzel, 2020; Tsui, 2011). Overall, researchers and practitioners have used

the zone of proximal development as providing a theoretical basis for grouping students according to higher versus lower achievement categories. By contrast, an alternative recommendation could be to make sure that each author receives feedback from at least one high achieving student.

However, there are often wide differences in students' knowledge and skills. A two-level higher versus lower achievement distinction, such as the expert versus novice distinction, can be a very gross approximation of the full range of achievement differences within a course, especially large courses. Both expert companions and novice companions can be further divided, such as dividing relative experts into highest versus higher achievement and relative novices into lowest achievement and lower achievement groupings. From a zone of proximal development perspective, optimal zones of feedback are likely narrower than a simple higher/lower division. On the other hand, precise matching to the exact same achievement level is unlikely to be feasible nor necessary.

General pros and cons of achievement grouping

As a way of within-class grouping, achievement grouping is defined as forming groups according to students' achievement differences within a specific or individual class (Hattie, 2009; Hollifield, 1987). Unlike ability grouping that is conceptualized as more fixed over time, achievement grouping is conceptualized as being more fluid and based upon specific tasks (Hattie, 2009). A number of positive aspects to achievement grouping have been identified beyond the context of peer feedback. As a way of supporting interactive learning, grouping students by achievement level allows different students to have greater freedom to choose learning materials to suit their group's learning speed, and further improve students' task performance (Buttaro & Catsambis, 2019; Francis et al., 2019). Grouping can improve friendships in a class (Kim et al., 2020). Grouping can also help students who lack motivation or interest in subject learning and provide them with additional support (McKeen, 2019). Students struggling in a particular subject area will not feel pressure and anxiety from having to catch up with accomplished peers when they are grouped with like peers, and students with stronger achievement can also gain a sense of accomplishment through achievement grouping (Matthews et al., 2013). These arguments have been used to support the creation of online peer feedback environments that automatically group students by achievement (e.g., Abrache et al., 2021; Giannoukos et al., 2010).

However, a number of concerns about achievement grouping have also been identified in general learning research. Achievement groups can have more peer conflict (Kim et al., 2020). Slavin (1993) noted that grouping according to achievement visibly tags students in way that can limit their potential. Further, students in low-achievement groups may experience reduced learning opportunities and therefore show reduced progress compared with students in high-achievement groups (Buttaro & Catsambis, 2019). A number of studies have found that low-achievement students have been found to do better in heterogeneous groups than homogeneous achievement groups, whereas high-achievement students did equally well in the other grouping format (Kanika et al., 2022; Saleh et al., 2005). Low-achievement groups may not be able to spot problems in their peers' work or provide high-quality feedback (Ariawan, 2018). Finally, students in low-achievement groups may have lower academic engagement (Steenbergen-Hu et al., 2016).

There are also logistical challenges to achievement grouping. Teachers may not have enough information to accurately group students for a particular task; it should be their task-specific achievements rather than broader achievements/academic performance that

matter. As a dynamic set of achievements unfolding over time, teachers may struggle to keep re-grouping appropriately (Sheppard et al., 2018). A meta-analysis on within-class grouping shows that achievement grouping only has a slight advantage over nongrouping ($d = 0.17$; Hattie, 2009; Hattie, 2012), and the effect occurs predominantly in larger classes (Lou et al., 1996). Some have argued that inconsistent effects of grouping may be caused by inaccurate group assignment (Buttaro & Catsambis, 2019). On the other hand, new algorithms can be created to dynamically assess student performance and create appropriate groups in each assignment. A number of researchers have examined learner grouping using computer algorithms (e.g., Henry, 2013; Klein et al., 2009; Wang et al., 2007) to improve pairwise transactivity (Wen et al., 2017), tie strength within teams (Salehi & Bernstein, 2018), course commitment, and compatible schedules (Hastings et al., 2022).

Achievement grouping and peer review

Although within-class grouping has been extensively studied in general, the applications of those general findings to peer feedback, in particular, are still unclear. First, multipeer feedback lies somewhere between pairs (because the initial author–reviewer interaction is a dyadic interaction) and groups (authors and reviewers interact with more than one other student in a round of reviewing). Second, peer feedback, especially computer-mediated peer feedback, is also highly focused on peer interaction processes per se. That is, a group working on a mathematics task in class can choose to work in parallel or let one student carry the primary workload, but peer feedback requires interaction. On the other side, the interaction in peer feedback is relatively brief, and opportunities to negotiate shared meanings within received feedback can be limited. These differences can greatly alter the relative advances of homogeneous versus heterogeneous grouping.

In peer review, the feedback comments, rather than the ratings, are regarded as particularly important for improving task performance (Cho & MacArthur, 2011; Wooley, 2007). Providing feedback to peers or receiving feedback from peers as part of revision work has been shown to promote learning (e.g., Chang et al., 2021; Tsivitanidou et al., 2018; van Popta et al., 2017). More recent studies applying regression analyses have provided some insights as to why: only the more active learning activities (providing comments, acting on received comments) produce learning; just receiving comments or just reading other students' documents as models, per se, are passive learning activities that produce little observable changes in a student's own task performance (Wu & Schunn, 2023a). A number of studies have examined how the amount of feedback provided or received is influenced by the relative achievements of authors and reviewers, typically using a higher versus lower achievement dichotomy (Wu & Schunn, 2021b; Day et al., 2022; Hsia et al., 2016; Zhang & McEneaney, 2020).

From a receiving feedback perspective, higher achievement authors are reported to receive a similar amount of critical feedback from lower or higher achievement reviewers, whereas lower-achievement authors receive more critical feedback from higher-achievement reviewers (Patchan et al., 2018; Huisman et al., 2018). In terms of impact on student revisions, mixed findings have occurred. In one study, lower-achievement authors implemented more of the feedback they received in their revisions when the feedback came from higher-achievement reviewers (Sunahase et al., 2019), but other studies found lower-achievement authors implemented more of the feedback from lower rather than higher-achievement peers (Huisman et al., 2018, 2019). Further, revisions were of a higher quality when based upon feedback from students of a similar achievement level than when based

upon feedback from students of a different achievement for both higher- and lower-achievement students (Ariawan, 2018; Patchan et al., 2018). Thus, from a receiving feedback perspective, there is mixed support for same achievement grouping.

From a feedback providing perspective, a number of overall achievement effects have been documented: lower-achievement reviewers provide more praise while higher-achievement reviewers provide more criticism (Cho & Cho, 2011). Having engaged deeply with the reviewed documents through providing comments, the reviewers with low achievements imitated the reviewed tasks as though they were models, while the reviewers with high achievements reconstructed and applied knowledge by providing suggestions, which shapes the learning opportunities of the reviewers. There has also been an interaction between reviewer and author achievement, higher-achievement reviewers note more issues in lower quality documents than higher-quality documents, while lower-achievement reviewers tend not to notice more problems in documents from authors with different achievements (Patchan & Schunn, 2015).

No studies to date have taken a finer-grained examination of the effects of relative achievement on the impact of providing or receiving peer feedback, likely because of concerns about statistical power (i.e., having enough students for each specific matched and mismatched achievement groupings): Two achievement levels produce 4 groups (high–high, low–high, high–low, low–low), three achievement levels produce 9 groups, and four achievement levels produce 16 groups. Unfortunately, the binary split (i.e., higher versus lower) on student achievement could significantly misrepresent the benefits of mismatching. Student performance within a class is often normally distributed, meaning most students are near the midpoint. Thus, many of the pairings of "higher" and "lower" achievement students might actually involve two students of very similar achievements (i.e., close to an average level). Second, small versus moderate versus large differences in achievement may matter. For example, when students as reviewers are only moderately higher, they may focus on useful issues and provide clear guidance; but when they are much higher, students may focus on issues that are "beyond" the zone of proximal development for the feedback recipient. In the current era where peer assessment is taking place with increasing frequency due to the growth of online peer assessment systems, it is now possible to conduct more fine-grained analyses with sufficient statistical power afforded by large datasets.

In addition, as mentioned earlier, although scholars have conducted extensive research on the general effect of peer feedback on changes in task performance, no research has examined the impact of the characteristics of peer feedback (e.g., number, length, quality) on changes in task performance with more fine-grained ability matching in peer review. In theory, receiving more comments leads to more revisions and more revisions lead to greater learning (Wu & Schunn, 2020), but not all peer feedback assignments require students to revise. Similarly, providing more comments is associated with greater learning gains (Zong et al., 2021; Cho & Cho, 2011). However, the relative learning benefit of each comment provided could vary by relative achievement differences. For example, providing feedback to lower achievement peers could simply involve practicing well-honed skills (i.e., contain relatively low benefits) whereas providing feedback to higher-achievement peers could involve extensive planning and thought. However, providing feedback to much higher-achievement peers could involve giving incorrect advice.

Conceptually, we extend the notion of a zone of proximal development to make an initial similar prediction of approximate match effect in terms of four achievement levels, which we argue captures noticeably different performance levels in larger classes without having to examine very small performance differences (e.g., the difference

between a student receiving a 78 versus a student receiving a 79 on an assessment). That is, the benefit of providing or receiving feedback will be best when students are at the same level within a four-achievement level distinction. Figure 1 shows the predicted changes in students' subsequent task performance for providing and receiving under this "match" hypothesis. However, the mixed prior findings in the literature suggest actual results are likely to be more complicated. First, adjacent achievement groupings may also be beneficial. Second, the benefits of providing are likely to be different from the benefits of receiving, both overall and from a matching/mismatching perspective. Note that changing task performance from one assignment to the next is a slow process given the wide range of skills, knowledge and dispositions students are trying to master, and that transfer of gains in skills, knowledge, and dispositions from one learning task (i.e., just receiving or just providing feedback) to improved performance in a later task are often small for a number of reasons. Thus, the relative impacts of matching within one assignment are likely to be small, and only grow in importance when accumulated across assignments.

Study aims

Based on critical gaps in the literature on peer feedback, this study explores major open questions about whether and why achievement pairings influence growth in students' performance across assignments through a combination of disciplinary learning and motivational changes. In particular, the study asks three exploratory research questions about the impact on growth in task performance across assignments of reviewer–author pairings using four levels of task performance:

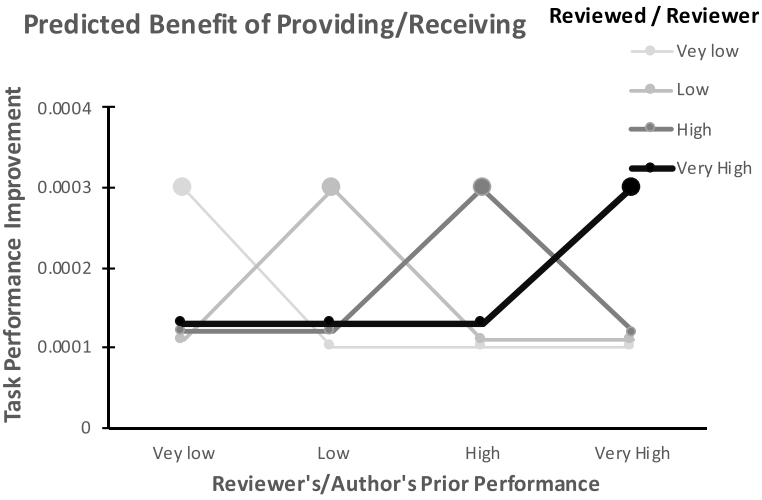

Fig. 1 According to a strong matching hypothesis, the predicted marginal means for performance benefit as a function for the level of reviewers' prior performance and authors' prior performance when considering four achievement levels

RQ1: To what extent does amount of change in task performance across assignments from providing feedback and of receiving feedback vary by pairings using four achievement levels?
RQ2: To what extent does the amount of feedback provided/received vary by pairings using four achievement levels?
RQ3: Taking into account both providing and receiving feedback effects, what are the cumulative effects on gains in task performance across a number of assignments of using an algorithm that matches students using four performance levels (reassessed within each assignment) rather than being randomly assigned?

Methods

Course setting and participants

Peer feedback and task performance gains across assignments was examined for 766 undergraduate students, the enrollees of three large biology courses ($n = 274$, 296, and 196, respectively) at two public research-oriented universities in the USA. These larger enrolment courses all made extensive use of peer assessment across the semester, collectively producing a sufficiently large amount of data to support statistical analyses of the benefits of participating in peer feedback within a variety of author–reviewer pairings using four student performance levels. Students rarely provided demographic details since that step was optional and entirely avoided when the instructor had accounts automatically created in the system by linking to the university learning management system. Publicly listed enrolment information for undergraduates at the two universities was as follows. The first university, the site of the first course, had undergraduate enrolment of approximately 55,000, who were 58% White, 25% Latinx, 9% Asian, 2% Black, and 3% two or more races. The second university, the site of the second and third courses, had an undergraduate enrolment of approximately 30,000, who were 44% Asian, 30% Latinx, 15% White, 2% Black, and 6% two or more races.

All the courses also used the same online peer assessment system, Peerceptiv, which collected all study data. Each course had four or five biology-related assignments of moderate difficulty involving analyzing research articles or writing up plans for experiments or results of studies, orchestrated to develop students' scientific analysis and communication skills. To complete a given assignment, students were required to participate in each stage of a peer review process: (1) completing and submitting task documents to the online system for peer review; (2) reviewing four peers' submissions and providing ratings and comments; and (3) rating the helpfulness of all feedback received to encourage higher-quality reviewing. This process was repeated across the four or five assignments in the course. Figure 2 shows a simplified representation of the multipeer reviewing situation and the way that it unfolded across assignments. In assignments, students at varying performance levels (different circle thicknesses) are assigned to provide feedback to students at varying performance levels in a random way, and the amount of feedback provided to/received from different students varies (different line thicknesses). Across assignments, students can improve or decline in task performance level and will thus be assigned different students to review. The diagram is simplified in terms of (1) only showing fours students in the class rather than hundreds and (2) only showing two reviews received/provided in each assignment rather than four.

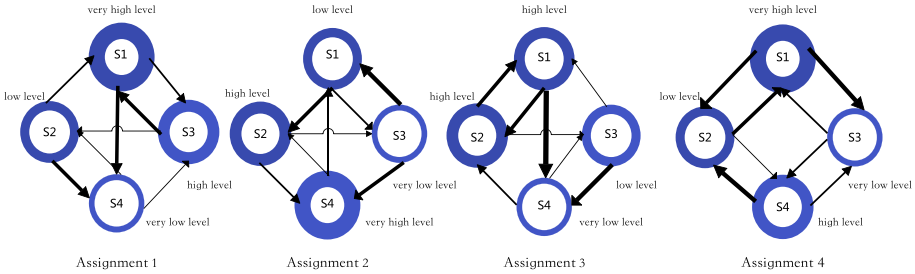

Fig. 2 A conceptual diagram showing the evolution of multipeer reviewing across four assignments in which students provide and receive to students at different current performance levels and then provide/receive varying amounts of feedback in those assigned reviewing pairs

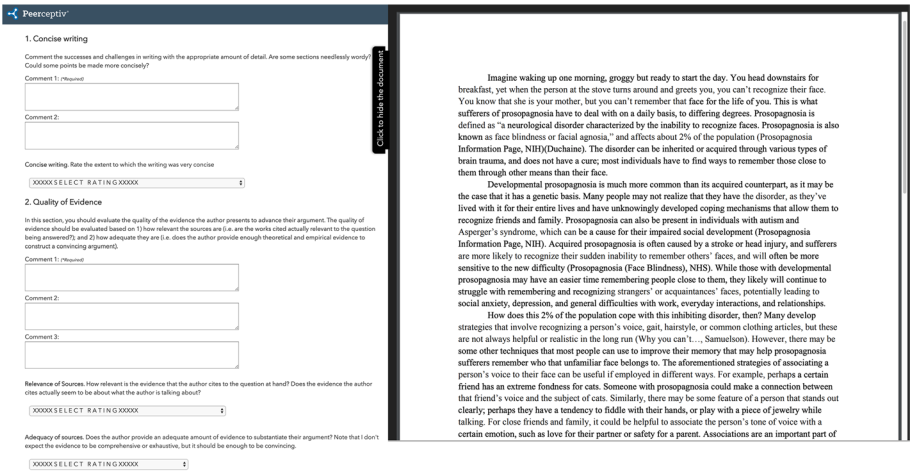

Fig. 3 The reviewing interface within Peerceptiv at the time of this study. A pdf document viewer is on the right, and interleaved comment textboxes and pulldown rating menus are on the left

Figure 3 shows the main student reviewing interface, which gave guidance on the comments provided as well as a number of analytic rubrics for providing ratings. Students needed to give at least one comment for each aspect of the reviewing task (e.g., concise writing, quality of evidence), there were no requirements for the number of words in a given comment.

Measures

The student reviewing behaviors (ratings and comments) were used as the basis for all the measures used in the current study (see Table 1 for an overview). Each measure was based upon the behaviors within a particular reviewing assignment, so that the reviewing experiences from one round could be associated with changes in task performance (the main dependent variable) from one assignment to the next assignment. In other words, the analytic approach involved a time-series analysis. Independent variables involved the amount

Table 1 Quantitative measurement variables for each construct and their definitions

Construct	Measure	Description
Task performance	*Z-score*$_J$	A student's task score on the *J*th assignment (i.e., mean ratings across raters and rating rubrics), standardized within that assignment (mean = 0, SD = 1)
The length of feedback comments provided	*Length provided*$_J$	The total number of words provided on the *J*th assignment
	Very-low length provided$_J$	The total number of words provided on the *J*th assignment to students with very low performance
	Low length provided$_J$	…to students with low performance
	High length provided$_J$	…to students with high performance
	Very-high length provided$_J$	…to students with very high performance
The length of feedback comments received	*Length received*$_J$	The total number of words received on the *J*th assignment
	Very-low length received$_J$	The total number of words received on the *J*th assignment from students with very low performance
	Low length received$_J$	…from students with low performance
	High length received$_J$	…from students with high performance
	Very-high length received$_J$	…from students with very high performance
Round	*Round*	The sequential assignment number in a course
Course	*Course*	Categorical dummy indicator (1 for first course, 2 for second, and 3 for third)

of feedback provided and received in the prior assignment, either overall or to/from students at particular task performance levels.

To match what would typically happen in a computer environment implementing a matching algorithm, task performance in the prior assignment was used to categorize students (as receivers and as a providers) in terms of relative achievements into four performance levels (very low, low, high, very high) at the time of the reviewing task. Four performance levels were used considering measurement precision (current student performance is measured with some noise and is only an approximately of underlying skills) and the idea of a zone in the zone of proximal development theory. Note that students themselves could change categories from assignment to the next as their task performance levels changed.

Procedurally, learning is inherently a complex construct to measure and typically requires some mathematical operations to approximate it. Some researchers measure learning in terms of a single assessment, but this approach assumes that performance was at chance or equivalent across learners' pre-tests. More typically, there is a pre–post change score to measure learning, but this change score approach requires relatively strict measurement equivalence between pre- and post-test. Another approach uses a regression that predicts post-test while controlling for pre-test, which does not require a strict measurement equivalence. Our time-series analysis is conceptually similar to the regression-predicting-post-while-controlling-for-pre approach. The approach, however, fails when the pre-test is a poor predictor of post-test. In our case, prior assignment performance generally was a good predictor of performance on the subsequent assignment, providing an important check of the underlying assumptions of the analytic approach.

A question that arises in all measurement approaches is whether there was knowledge gain/skill development or a motivational change. We argue that a motivational change is important and falls under broader definitions of learning, unless the motivational change is a relatively temporary change, like situational interest (Hidi & Renninger, 2006) or task achievement goal (Urdan & Kaplan, 2020). A recent meta-analysis revealed that peer reviewing can improve student motivation, such as self-efficacy, in addition showing gains in performance assessments (Li et al., 2021). Since there were weeks between assignments in the studied courses, if the basis of the observed relationships in the current study were motivational, it was not a temporary motivational change.

Another question that arises is whether the underlying changes driving task performance differences are relatively narrow in scope (such as facts in skills closely related to the assignment/assessment) or broader in scope, such as a general trait, for example, brain training in various forms, targeting working memory, attentional control, or frontal lobe development (Gobet & Sala, 2023). However, meta-analyses suggest such broad ability changes are weak, and they always require very intense activity across many sessions (Sala & Gobet, 2020a, b). Therefore, the experiences from peer reviewing a few peers' documents are unlikely to reflect general ability changes and we assume instead observed changes in task performance reflect changes in discipline and genre-specific knowledge, skills, and attitudes (e.g., related to research writing in biology).

Task performance In these three courses, task performance was evaluated by four peers using detailed rubrics that had five to six different ratings, which are easily and automatically obtained performance estimates for large numbers of students across many assignments. Obtaining expert ratings for so many students across so many assignments would be very difficult to obtain, and even automated essay scoring systems would require many expert scores to support their training across so many different writing assignments. The large number of

peer ratings per document (four peers multiplied by four to six ratings dimensions per review) produced an overall score for the task. In each assignment, students are randomly assigned to a new set of peer reviewers, so the relative matching status from the prior assignment (the IV) would not influence measurement characteristics of the outcome variable, which is performance in the next assignment. Further, more recent studies show little relationship between students' own performance and the accuracy of their reviews when given a good rubric (Xiong & Schunn, 2021, Wu & Schunn, 2023b). Finally, the use of the regression approach rather than a change-score approach to measuring task performance improvements means that small systematic bias in prior assignment grades would have relatively little influence on the regression outcomes. A number of studies, including a meta-analysis (Li et al., 2016), have established that such an approach (online multipeer assessment using well-structured, multipart rubrics) typically produces valid and reliable mean ratings, often at the same level of validity as trained expert raters. With the accountability mechanisms which produce higher levels of validity in Peerceptiv (accuracy grades for consistent scoring and helpfulness grades for comments perceived to be useful), reliability and validity even at the individual rating dimensions is typically adequate (e.g., ICCs and validity correlations averaging around 0.45 across thousands of rubrics; Wu & Schunn, 2021b).

To nuisance variance due to differential complexity of each assignment and changes to aspects of the rubrics (e.g., adding or deleting rubric dimensions; changing level expectations within a rubric dimension), we calculated a $Z\text{-score}_J$, a standardized score of the task on that Jth assignment within a course (total task score minus assignment-specific mean and then divided by the assignment-specific standard deviation).

The length of feedback comments provided Within the peer review system, students were required to provide comments to each peer along a number of reviewing prompts. But the structure of the comments varied widely, including aspects studied by prior research: namely, simple versus elaborated (Strijbos & Sluijsmans, 2010), general versus specific solutions, and descriptions versus explanations (Cho & MacArthur, 2011; Tseng & Tsai, 2007). However, such variations in content will likely be correlated with the length of comment (e.g., Patchan et al., 2018, Wu & Schunn, 2021b) such that longer comments likely had more features that make comments useful. In addition, students sometimes entered multiple comments in one comment box or separated comments across multiple comment boxes. Therefore, we defined the measure to be based upon the total number of words summed across comment boxes rather than number of comments or mean number of comments per box.

Further, it was expected that the total length of comments provided by students in a particular review would vary depending on the quality of the task reviewed and the students' ability to detect problems related to each rubric dimension, provide possible solutions, and explain their reasoning. In other words, the pairings of student author/reviewer according to their task performance level was expected to influence length provided/received, which then is expected to shape the learning opportunity.

A prior study applying meta-regression to data from 2421 students across 13 courses revealed that the length of comments provided (total number of words provided across reviews) in particular is a highly consistent and powerful predictor across course contexts of growth in academic performance across the semester, even when including a wide range of controls (Zong et al., 2021). As a result, the contents of peer feedback (provided and received) are viewed as the main source of learning. Further, to ground the simulation we created in the second part of this study to test the predicted effects of matching reviewers across assignments, it was important to estimate the specific per-word benefits of providing and receiving.

For the analyses of the effects of feedback provided within different author–reviewer performance pairings, a four-level achievement distinction was used to capture the amount of feedback provided to authors at a given level. In particular, we classified the students according to their different task performance on the prior assignment and then calculated the length of feedback provided by specific learners to authors at different performance levels: the length of feedback comments provided to students at each of four different performance levels: very low, low, high, and very high performance (using quartiles applied to the *Z*-score task performance measure). The amount provided by a reviewer to a particular author achievement level was a function of how much reviews the reviewer was assigned, how many they actually completed, the number of problems that existed in each reviewed document, and the reviewer's tendency to provide in-depth reviews.

The length of feedback comments received Not surprisingly, peer feedback received also influences student performance, and prior research has established that the total length of feedback received from peers is often a predictor of gains in students' performance, although often a weaker predictor than feedback provided and sometimes a negative predictor (Zong et al., 2021), presumably because receiving a large amount of critical feedback can be demotivating. Similar to the feedback provided, we calculated four measures that involved the length of feedback received from students of different performance levels on the prior assignment, again using the same four performance levels (very high, high, low, and very low).

Comment features To provide qualitative data to deepen the findings of this primarily quantitative investigation, ten randomly selected long feedback comments were selected for each of the four extreme group achievement pairings (very low and very high reviewers matched to very low and very high authors) for text analysis. Focusing on long comments (defined as length of having at least 40 words) enabled examination of whether more extended comments were different in nature depending upon the achievement pairing. In particular, these selected long comments were coded for four different components that they could have contained: criticism, suggestion for revision, explanation of issues, and praise. The specific components have been regularly examined in resource of peer feedback (e.g., Patchan et al., 2018; Huisman et al., 2017) and have been implicated in influence revision behavior and changes in task performance of providers (Zong et al., 2021). Some reviews have multiple or even all features, while others may have only two or even one feature (see Table 2 for examples of each feature).

Analyses

The primary analysis approach involved multiple regression to examine the extent to which the reviewing experiences in one assignment (i.e., amount of provided and received comments) predicts improvements in task performance in the next assignment. It has already been established across a wide range of courses that the length of comments provided and received are robustly the best predictors of improvements in task performance in the next assignment (Zong et al., 2021). At issue in the current set of analysis is the relative contributions of reviewing experiences under different author–reviewer performance pairings. Four separate regressions focus on the level of the author themselves (i.e., one regression for authors who were very low in the prior assignment, a second regression for authors who were low in the prior assignment, etc.), and then four different predictors consider either the level of the provider of the comments (for the Received variables) or the levels

Table 2 Peer feedback coding scheme. Italicized text highlights the aspects of the examples that best matched the feedback code

Type of feedback	Example
Criticism	*I feel your lay summary was too vague*. Although this is supposed to be written for a lay audience, it does not mean that they are incapable of understanding scientific material. *I feel you went too broad and should have included more detail*
Suggestion	*One way to perhaps address this issue is to think about the following questions: what is the purpose of understanding sleep drive? Why does the lay audience need to be aware of these major concepts? If the writer is explaining the paper for the first time to a friend, how would the writer breakdown scientific information to keep his or her friend interested and scientifically aware?*
Explanation	*I gave this paper this grade because I think that it shows deep understanding of the material presented*. It showed great transition from general topic to focused background topic and was great in elaborating the data from the studies
Praise	*I really liked and was inspired by your second paragraph. I always wanted to do something similar to you and your plan gave more ideas and inspired me for my future. I also really liked your idea regarding the medical camp and I also wanted to do something like this as well*

of the recipient of the comments (for the Provided variables). Note that separate models were used to estimate level-specific providing and receiving effects to avoid unstable models with too many simultaneous predictors. However, every model had both providing and received feedback predictors—what was varied was the level of detail in each predictor.

The dependent variable, the standardized task performance in a given task ($Z\text{-score}_J$), is a continuous variable with normal distribution. Therefore, linear regression models were used, with the $Z\text{-score}_J$ as the dependent variable, $Z\text{-score}_{J\text{-}1}$ as its baseline in the time-series model, and then very-low length provided/received$_{J-1}$, low length provided/received$_{J-1}$, high length provided/received$_{J-1}$, very-high length provided/received$_{J-1}$ as the core predictors, and total length provided/received$_{J-1}$, round, as additional control variables. Dummy codes for each course were also included in the regression. The Stata code implementing these regressions is presented in Appendix C. Appendix Table 5 shows the mean and standard deviations for each variable. No predictor had restricted range issues. Appendix Tables 6, 7 shows the correlations among predictors, separately by each author performance level, since separate regressions were conducted for each author performance level. Variance inflation factors were examined to address concerns about multicollinearity, and no predictors showed concerning values (all VIFs < 2.0).

To follow-up the quantitative findings, we then examined patterns in the comment features that tended to be found in the long comments of various reviewer pairings. For example, did changes in comment features provide potential explanations for differences in benefits of providing or receiving in various achievement pairings?

The results of the multiple regression analyses described above tested the relative contribution to task performance growth per word provided (or per word received), ignoring how many words are typically produced in each achievement pairing; this tests whether the contents of the comments provide a learning opportunity in the zone of the learner. But the number of words provided and received were also expected to vary substantially according to the performance levels of authors and reviewers (i.e., lower performing authors were expected to receive more comments and higher performing reviewers were expected to provide more comments). Thus, the net effect of author–reviewing matching is also expected

to be importantly modulated by the amount of feedback provided/received rather than just being placed in a given experiential context. We therefore also calculated how many words are produced on average in each of the 16 author–reviewer achievement pairings to provide a more complete picture about the learning opportunities by author-reviewer pairing. A two-way repeated measures ANOVA (four author levels × four reviewer levels) was applied to the number of words found in a review to test the statistical significance of observed patterns.

Simulation

Finally, based on the estimated effect values obtained from empirical results described above (i.e., per word benefit of providing/receiving and amount of provided/received feedback within each of the 16 author–reviewer performance level combinations), we conducted a simulation implemented in Python 3.7 to make predictions for overall growth across multiple peer reviewing assignments in students with different initial performance levels (very low, low, high, very high) as a function of two different methods for assigning students to reviewing tasks (random assignment and matching performance levels) to test whether students would experience greater overall growth in task performance if consistently matched with peers at their same level. In particular, the simulation provides an integrative understanding of the joint effects of reviewer assignment matching on (1) learning from receiving and learning from providing comments, (2) opportunity to learning from amount of comment words received and provided, and (3) the cumulative effects from being consistently matched across multiple assignments.

As an input to the simulation, the estimated per-word effects (as unstandardized betas) were multiplied by the estimated mean number of words provided/received to produce estimated effects for providing in one assignment and receiving in one assignment. Then, the cumulative effects from both providing and receiving across five assignments (each assignment involving students being assigned to four peer reviews) were calculated using the simulation. In the case of the random assignment algorithm, the simulation was run 1000 times since each run involves noise due to the inherent randomness of this approach. The simulation code is included as an online attachment.

Results

Per-word learning benefits from providing and receiving feedback

Table 3 presents the standardized betas from the multiple regression models estimating the benefits of providing feedback to students in each performance level, separately for students with different starting performance levels, controlling for differences in amount of received feedback. Note that this regressions measures "transfer" of experiences in reviewing in just one assignment to growth in task performance in the next assignment, and they test the separable benefits from minor variations of the same learning opportunity (learning from providing) in a multiple experience content. Thus, it is to be expected that the effect sizes of each predictor are relatively small, given the challenges of finding transfer and separating out the growth into related competing predictors. These small individual effects nonetheless can accumulate to pragmatically meaningful effects across assignments, as will be tested by the simulation. Further, of particular interest is the pattern of effects across all the 16 author–reviewer pairings rather than the statistical significance of any one regression coefficient, therefore, post hoc corrections for number of statistical tests were not applied.

Table 3 For students reviewers starting at the four different relative performance levels in the prior assignment, standardized betas from the multiple regressions predicting improvements in assignment performance as a function of providing comments to authors at varying performance levels, along with regression *N*

Predictor	Relative performance group of reviewers in prior assignment			
	Very low	Low	High	Very high
Baseline				
Z-score$_{J-1}$	0.07*	0.06*	0.01	0.11***
Core predictors (to whom provided)				
Very-low length provided$_{J-1}$	0.11**	0.06	0.03	0.10**
Low length provided$_{J-1}$	0.01	0.07*	0.06	0.05
High length provided$_{J-1}$	0.07*	0.07*	0.08*	−0.001
Very-high length provided$_{J-1}$	0.08*	0.08*	0.04	0.06^{t}
Control variables				
Total length received$_{J-1}$	−0.19***	−0.01	−0.05	0.07
Round	0.06*	−0.01	0.07*	−0.04
Course (versus Course 1)				
Course 2	−0.09	0.13*	0.05	0.18***
Course 3	0.02	0.17*	0.01	0.11^{t}
N	937	963	945	957
R^2	0.05	0.02	0.02	0.06

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $^{t}p < 0.1$

In all cases, providing feedback to students at the same performance level appears to be beneficial. In addition, for the reviewers with very-low or low prior performance, three of the four core predictors are significant (i.e., the lower performing students generally benefit from providing feedback to others). For the reviewers with higher prior performance, the benefits of reviewing are more restricted. In the case of high prior performance reviewers, only providing to same achievement peers was statistically significant. In the case of very-high prior performance reviewers, providing benefits were seen for both same achievement (low) and opposite achievement (very low). It is also interesting that total number of received comments had a sizeable negative effect, but only for the very-low reviewers, potentially speaking to a negative motivational effect to such students who presumably received a lot of criticism and relatively little praise.

Table 4 presents the standardized betas from the multiple regression models in which students within different prior assignment performance groups received feedback from students in each prior performance level, controlling for differences in total amount of provided feedback. Although there continued to be evidence of strong overall benefits of providing feedback (the cumulative effects of providing to students at various levels), the general effects of receiving feedback were rarely statistically significant, and the effects were negative when the effects were statistically significant for particular matching groups. As a pattern, the lower the prior performance group of the feedback receiver, the more subgroups of feedback providers showed negative effects and the larger the negative effects on the feedback receiver. There was no specific pattern in which same achievement pairings were a particularly bad choice.

The variation in standardized betas by student performance levels presented in Tables 3 and 4 could potentially be explained by variation in the relative amounts of feedback

Table 4 For student authors starting at the four different relative performance levels in the prior assignment, standardized beta coefficients from the multiple regressions predicting improvements in assignment performance as a function of receiving comments from reviewers at varying performance levels along with regression *N*

Predictor	Relative performance group of authors in prior assignment			
	Very low	Low	High	Very high
Baseline				
Z-score$_{J-1}$	0.08*	0.07*	0.01	0.11***
Core predictors (from whom received)				
Very-low length received$_{J-1}$	−0.09*	−0.07*	−0.01	−0.03
Low length received$_{J-1}$	−0.04	−0.01	−0.04	−0.03
High length received$_{J-1}$	−0.09*	−0.07*	−0.02	0.02
Very-high length received$_{J-1}$	−0.09*	0.06	−0.03	0.05
Control Variables				
Total length provided$_{J-1}$	0.20***	0.20***	0.13**	0.18***
Round	0.05	−0.01	0.07*	−0.13***
Course (versus Course 1)				
Course 2	−0.09	0.11*	0.05	0.18**
Course 3	0.02	0.14*	0.01	0.11*
N	937	963	945	957
R^2	0.05	0.03	0.02	0.06

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

provided/received (i.e., from restricted range issues) or in the benefits of feedback provided/received. To isolate the effect components, we next examined the unstandardized regression coefficients from the same regressions (i.e., the per-word benefit of comments provided and received; see Fig. 4). For the per-word effect of providing feedback (Fig. 4 top), the marginal benefits of reviewers providing feedback to students at the same level (indicated with large circles) are relatively higher than those of students at other levels. This matching performance level effect is particularly large for very-low reviewers. However, for students with very-high prior performance, providing feedback to students with very-high performance appears to be less beneficial than providing to students with very-low or low performance. Trending in the same direction, high reviewers also appear to benefit from providing to low and very-low authors. Finally, the variation in amount of benefit is substantial across the groupings: the highest per-word benefit is roughly four to five times the lowest per-word benefit. Note that the "anti-match" cases (i.e., extreme mismatches; very-low students providing feedback to very-high students or vice versa) were not especially bad; the specific worst cases followed a more complex pattern.

Turning to the per-word benefit of receiving, we again see that receiving more feedback often has a negative effect on future performance, with only three exceptions: high or very-high authors experienced benefits from receiving more feedback from very-high reviewers, and very-high authors experience benefits from receiving more feedback from low reviewers). There is also one salient main effect: receiving from very-low reviewers generally has the largest negative per-word effect. This general pattern suggests the negative effect of receiving is unlikely to only be a negative reaction to receiving too much criticism because the very-low reviewers are unlikely to be the ones that provided the most critical long feedback. Instead, it may also be the case that

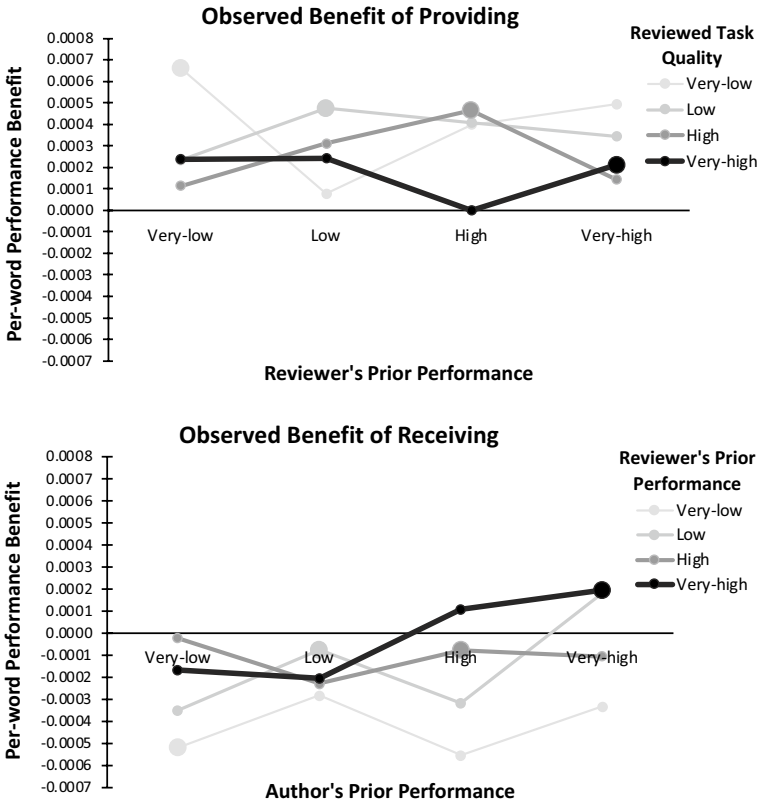

Fig. 4 The estimated per-word benefit of feedback provided (top) and received (bottom) as a function of reviewers' prior performance and authors prior performance. Matching achievement cases are shown with larger dots

receiving too much praise reduces motivation to improve. Instead, it may be that feedback needs to be balanced to consistently lead to improvements. Alternatively, it may be that very-low reviewers are less likely to provide suggestions for how to improve or they may provide explanations and suggestions that are not useful for learning. These alternatives are examined in further depth in the next section.

Qualitative differences in comments provided

To understand whether the contents of long comments differed across pairings, we examined variation in the contents of long comments in the four extreme group pairings: very low to very low, very high to very high, very low to very high, and very high to very low. Figure 5 shows that long comments almost always contain three features regardless of pairing: criticism, suggestion, and explanation. The only thing that is less routinely found in long comments and varies substantially by pairing is praise: praise occurred in long comments less than half of the time. Further, as one would expect, compared with very-low

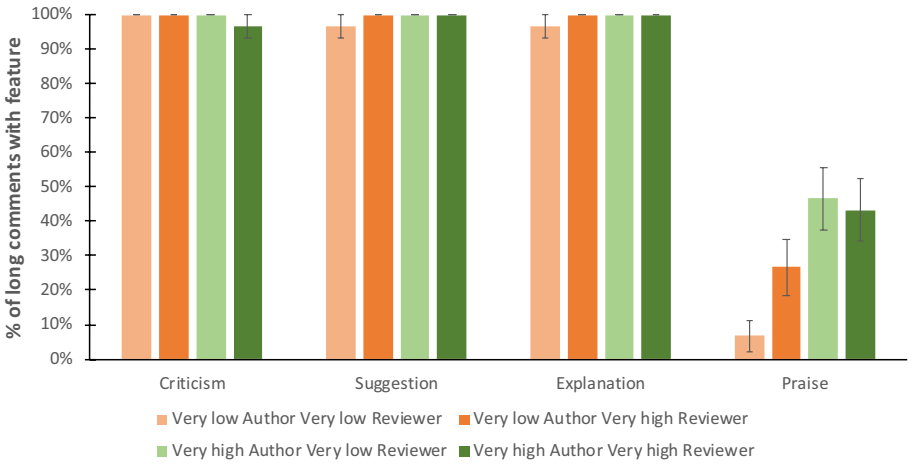

Fig. 5 The percentage of long comments (with SE bars) containing each feature in the various extreme author–reviewer performance pairings

performance authors, very-high performance authors were more likely to get praise (and equally so from very-low and very-high reviewers). Thus, the generally substantial negative effect of receiving more feedback from very-low reviewers is unlikely the result of too little criticism or too much praise. Further, since the benefits of providing feedback have previously been established to reside predominantly from constructive criticism with supporting explanations (Wu & Schunn, 2021b), the variation in degree of benefit from providing across the various author–reviewer matchings are unlikely to stem from the form of the content provided.

Amount of feedback provided and received

The prior analyses examined the per-word effects and contents of long comments. Another critical variable in understanding benefits is variation in the amount of feedback provided and received as a function of author–reviewer pairings. Figure 6 shows the mean total number of works provided/received within each pairing per comment dimension inside a review; note that providing and receiving are collapsed here because the amount provided is necessarily equal to the amount received within a student performance level pairing. Quantitatively, the results of repeated measures ANOVA found only two statistically significant main effects. First, there was a large effect of reviewer performance level ($\eta^2 = 0.11$, $p < 0.001$): the higher the reviewer performance level, the longer the feedback provided (mean number of words: very low = 44; low = 47; high = 50; very high = 55). Second, there was a moderate effect of author performance level ($\eta^2 = 0.06$, $p < 0.01$): the lower the quality of the document being reviewed, the longer the feedback it received (means: very low = 54; low = 49; high = 47; very high = 47). The interaction of reviewer performance by author performance was very small and not statistically significant ($\eta^2 = 0.01$, $p > 0.9$). These two main effects on the amount provided did not correspond to the variation in effects in the per-word benefits, suggesting that restricted range or other explanations of amount provided could not explain the variation in per-word benefits.

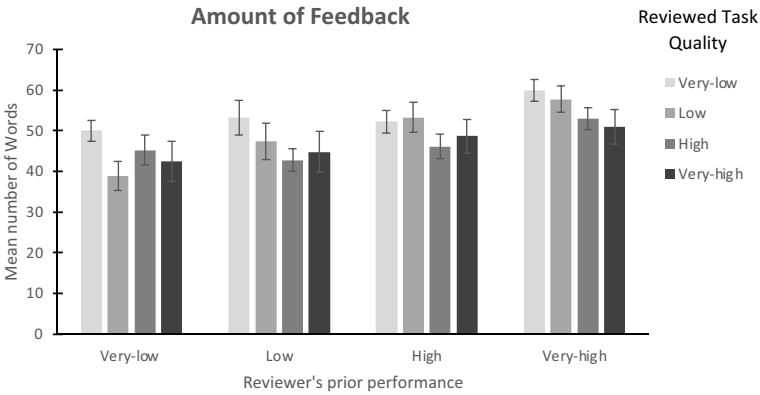

Fig. 6 The mean (with SE bars) number of words provided/received per assignment as a function for the level of reviewer's prior performance and the reviewed task quality

Simulation results

Integrating the per-word effects and total number of words produced/received estimated in the prior analyses, the results of the simulation (see Fig. 7) revealed the expected changes in performance for students starting at different levels as a result of experiencing one of two different reviewer assignment methods (matched or random) across five assignments with peer reviewing. The findings of the simulation are presented as estimated effects from providing, estimated effects from receiving, and total effects from the combination of providing and receiving.

As an overall orientation, the estimated providing effects (thick solid lines) were all various degrees of improvements in task performance, and estimated receiving effects (dotted lines), with one exception, were all various degrees of reductions in task performance. The combined effects of both providing and receiving (lines with squares) were always small positives.

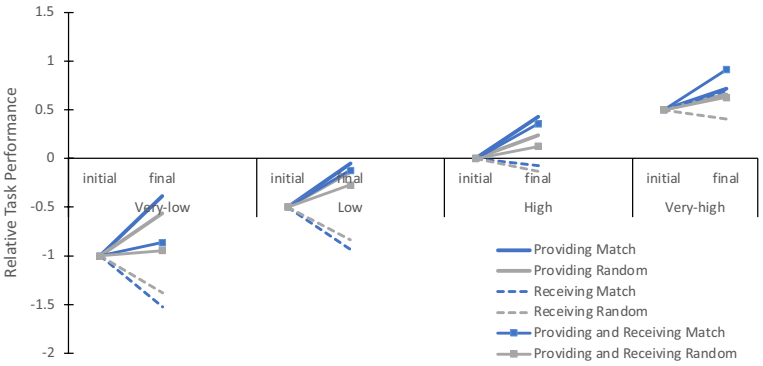

Fig. 7 The simulation's estimates for changes in task performance (initial performance to final performance after five peer review assignments) for students at four different initial performance levels and experiencing either performance matching or random matching, separating out the effect of receiving feedback alone, providing feedback alone, and the combined effects of receiving and providing feedback

From the receiving perspective, in two cases (very low and low), the negative effects were larger with matching, and in the other two cases (high and very high), the negative effects were weaker with matching or even more positive with matching. From the providing perspective, the matching reviewer assignment algorithm always produced larger estimated benefits than does the random assignment algorithm. When considering both receiving and providing effects, the combined effects always favored the matching algorithm. The difference between the algorithms at the level were relatively small (approximately 0.15 SD) for very-low and low students and moderate (approximately 0.25 SD) for high and very-high students. In other words, at least for high and very-high students, the cumulative effects of the matching algorithm were estimated to be pragmatically meaningful across assignments when considering both providing and receiving effects, despite the relatively small per-assignment effects of just providing or receiving effects, per word and total words, in isolation.

General discussion

The main aim of this study was to uncover the changes in student performance over time, associated with providing and receiving peer feedback within different prior student performance pairings when implemented as four performance levels rather than the typical high/low binary studied in the past. The study revealed important variations by four-level distinctions in student performance and generally supported the need to attend to author–reviewer relative performance levels within both providing and receiving feedback. Here, we consider the theoretical and practical implications of these findings.

Theoretical implications

For theories of computer-supported peer assessment, the current findings highlight importance of separate consideration of providing from receiving effects. In both the research literature and in the examined courses in the current study, providing feedback was associated with improvements in student performance. Prior research has long acknowledged that receiving is less beneficial than providing (e.g., Li et al., 2010; Lundstrom & Baker, 2009), but only relatively recently have negative impacts of receiving extensive feedback been acknowledged (Zong et al., 2021). Presumably very different mechanisms underlie these effects: providing influences performance gains via conceptual understanding and practice mechanisms (Wu & Schunn, 2023a; Chang et al., 2021), whereas receiving feedback may lead to reductions in performance via motivational mechanisms. The patterns of findings across performance pairings further highlight the importance of the receiving versus providing separation. For example, very-low performing students showed the highest gains from providing feedback in the exact match case but also the worst declines from receiving in the exact match case. If there was one underlying mechanism for both providing and receiving effects (e.g., practice opportunities), then there should have been more symmetry in the patterns within providing and receiving.

The particular patterns of relative gains and relative declines in performance associated with providing and receiving feedback now serve as a challenge for theories of peer feedback. There was some support for the general zone of proximal development theory

in that the exact match case was frequently more beneficial/less harmful. Furthermore, through the creation of groups by four levels of achievement, this study confirms that there is a stepped development zone in peer review, and students can gradually catch up to higher achievement peers with the help of peers with the same achievement. However, the support was far from a strong endorsement of that theory. Most critically, sometimes anti-match was actually more helpful than intermediate author–reviewer performance mismatches. Refinements to theories of learning from peer feedback are required to explain these patterns. Past studies using interviews and surveys with students suggested support for both learning from negative exemplars and positive exemplars during providing feedback (Schunn et al., 2016). Potentially, the variation in students' performance levels influenced their ability to notice positive and negative aspects of a peers' contribution (Huisman et al., 2018).

Better theoretical understanding of the basis of overall gains/declines from peer feedback experiences and relative gains/declines from peer feedback with different performance matchings is important to producing improvements in scaffolding with online peer feedback systems. For example, if the total amount of received feedback tends to overwhelm some learners, only a subset of the feedback could be presented. Alternatively, if some author–reviewer performance pairings lead reviewers to focus on problems outside of their zone of proximal development, rubrics within peer assessment could be dynamically adapted to the author-reviewer pairing.

Practical implications

At the broadest level, the findings of the current study providing additional support for the value of peer feedback for student performance outcomes. Meta-analyses have consistently found learning benefits of peer feedback in general (e.g., Li et al., 2020), and computer-supported peer feedback in particular (e.g., Chang et al., 2021). The current study adds to this research by suggesting that peer assessment is overall positive in its effects across student performance levels within a course. Regardless of where students started, both random-assignment and matching-assignment algorithms were associated with positive effects over time. This finding addresses student and instructor concerns that peer feedback helps lower-performing students at the cost of higher-performing students.

The results of the simulation also suggest that an achievement matching algorithm for reviewer assignment would generally lead to better outcomes than would the standard random-assignment algorithm. The benefit for matching over random was larger for higher-performing students but students in all four performance groupings showed the relative benefit of matching over random. Since the empirical patterns of providing, receiving, and amount of feedback were more complex than simply supporting a precise matching benefit, it is likely that more complex algorithms could be developed to further optimize outcomes.

Finally, the study pointed to the importance of long comments. Part of the value of computer-supported peer feedback is that mechanisms can be embedded to encourage students to provide longer comments: anonymity, grades for helpful feedback (Patchan et al., 2018), or automated feedback on the contents of a comment (e.g., Ramachandran, et al., 2017). The current findings suggest that lower performing students will especially be in need of support for producing longer feedback.

Limitations and future research

It is worth noting that the current study used correlational rather than experimental methods, so the causal relationship between experience and providing feedback is not fully established. However, prior experimental studies have already established the overall causal relationship between providing peer feedback and student outcomes (Li et al., 2020). Further, the regression approach did use a time-varying method and included many important control measures to address concerns about reverse causality and many third variable confounds. However, future research using experimental manipulations of different reviewer assignment algorithms will be helpful to fully establish the causal nature of relatively benefits/harms of different reviewer pairings.

It is also important to note the lack of measures of student motivation. The negative effects associated with receiving feedback were hypothesized to be caused by motivational mechanisms, but no direct assessment of changes in student motivation were included in the study. It is challenging to included multiple rounds of motivational assessments in the kinds of large courses needed to uncover effects of finer-grained author–reviewer pairings. However, the current study suggests that such an investigation will be needed to better understand how motivation is changing with peer feedback experiences.

Finally, the current study was done with university students in science courses using one online peer feedback system. The complexity of the assignments, the relative motivation levels of the students, and the supports included in the online peer feedback system could all shape the benefits of providing and receiving feedback, which could then shape the patterns of results in author–reviewer pairings. However, the study did find general consistency in the general pattern of results across the three courses that were examined.

Conclusions

The current study reveals that an approach to relative student performance levels using four achievements levels can inform the likely benefits to be obtained from peer feedback. In particular, providing feedback to students at the same performance level within four achievement levels appears to be especially helpful, but receiving feedback from students at the same level is not consistently better. Second, there appear to be simple main effects of both author performance levels and reviewer performance levels, rather than interactions involving relative match between the two, on the amount of feedback provided/received. Third, providing feedback, regardless of pairing, is associated with improvements in students' task performance to varying degrees, whereas receiving feedback, with only one exception, is associated with varying degrees of reductions in task performance. Finally, matching reviewers to authors by achievement level using four levels is expected to produced better outcomes in task improvement across courses than from randomly assigning reviewers, and more online peer assessment systems should explore adding matching algorithms as an option.

Appendix A

Table 5 Mean and standard deviations for each variable within each achievement pairing, along with maximum observed values on each variable

Variable	Very low			Low			High			Very high		
	Max	Mean	SD	Max	Mean	SD	Max	Mean	SD	Max	Mean	SD
Prior Z-score	−0.44	−1.37	1.04	0.20	−0.13	0.19	0.71	0.45	0.15	2.32	1.03	0.25
Very-low length Provided	1678	87.8	180.9	1935	124.9	244.0	3385	129.6	243.1	4994	178.6	402.1
Low length Provided	1445	56.2	192.2	1745	62.1	150.9	2017	82.1	180.7	1740	89.6	189.2
High length Provided	2131	84.8	193.2	1827	93.2	181.8	1833	91.4	171.9	2146	116.9	220.4
Very-high length Provided	1593	106.6	186.2	1768	130.3	225.2	2219	136.8	247.3	2309	144.3	247.8
Very-low length received	1654	90.1	181.4	1764	90.8	188.9	2736	96.7	205.2	1791	89.4	184.6
Low length received	1278	72.5	151.1	2105	70.8	169.3	1895	70.2	151.4	1249	64.1	135.4
High length received	1239	100.2	172.9	1858	99.8	199.2	1861	98.6	187.8	1561	80.1	163.6
Very-high length received	3343	187.3	310.9	2429	187.9	310.9	2330	162.2	261.9	1861	133.9	212.2
Total length Provided	4265	335.4	497.8	3926	410.6	542.8	4191	438.9	557.9	6772	528.9	745.9
Total length Received	3876	450.1	537.9	3537	449.4	554.7	3195	427.3	521.2	2516	367.4	448.6

Appendix B

Table 6 Within each performance level, Pearson intercorrelations among predictors and the outcome variable

	Z-score$_{J-1}$	*Very-low length provided*	*Low length provided*	*High length provided*	*Very-high length provided*	*Length received*	*Round*	*Course*
Very low Z-score$_{J-1}$								
Very-low length provided	0.003							
Low length provided	−0.03	0.27***						
High length provided	0.02	0.44***	0.30***					
Very-high length provided	0.04	0.30***	0.36***	0.42***				
Length received	0.01	0.50***	0.38***	0.47***	0.47***			
Round	0.01	−0.11	−0.05	−0.06*	−0.08*	−0.13***		
Course	−0.03	−0.44***	−0.37***	−0.41***	−0.45***	−0.71***	0.22***	
Z-score$_J$	0.08*	0.06*	0.02	0.07*	0.06*	−0.05*	−0.04	0.05
Low Z-score$_{J-1}$								
Very-low length provided	−0.03							
Low length provided	0.03	0.23***						
High length provided	−0.06*	0.28***	0.27***					
Very-high length provided	−0.07*	0.33***	0.19***	0.27***				
Length received	−0.15	0.50***	0.36***	0.46***	0.51***			
Round	0.04	−0.11***	−0.02	−0.07*	−0.09**	−0.13***		
Course	0.16	−0.43***	−0.34***	−0.43***	−0.43***	−0.72***	0.25***	
Z-score$_J$	0.08*	0.04	0.06	0.06	0.05	0.01	0.01	0.02
High Z-score$_{J-1}$								
Very-low length provided	−0.05							
Low length provided	−0.05	0.32***						
High length provided	−0.01	0.26***	0.14***					

Table 6 (continued)

	$Z\text{-score}_{J-1}$	*Very-low length provided*	*Low length provided*	*High length provided*	*Very-high length provided*	*Length received*	*Round*	*Course*
Very-high length provided	−0.003	0.28***	0.18***	0.24***				
Length received	−0.04	0.51***	0.39***	0.45***	0.47***			
Round	0.06*	−0.11***	−0.03	−0.10**	−0.10***	−0.14***		
Course	0.09**	−0.47***	−0.38***	−0.41***	−0.45***	−0.72***	0.26***	
$Z\text{-score}_{J}$	0.01	0.03	0.05	0.06*	0.03	0.01	0.07*	−0.02
Very high $Z\text{-score}_{J-1}$								
Very-low length provided	0.18***							
Low length provided	0.12***	0.34***						
High length provided	0.14***	0.29***	0.34***					
Very-high length provided	0.16***	0.32***	0.31***	0.24***				
Length received	0.11***	0.50***	0.42***	0.46***	0.50***			
Round	−0.09**	−0.05	0.02***	−0.06*	−0.06*	−0.09**		
Course	−0.17***	−0.42***	−0.38***	−0.43***	−0.43***	−0.53***	0.22***	
$Z\text{-score}_{J}$	0.15***	0.12***	0.08*	0.04	0.09**	0.06*	−0.12***	−0.09**

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

Table 7 In each performance, Pearson intercorrelations among predictors and the outcome variable

	Z-score$_{J-1}$	*Very-low length received*	*Low length received*	*High length received*	*Very-high length received*	*Length provided*	*Round*	*Course*
Very low Z-score$_{J-1}$								
Very-low length received	0.02							
Low length received	−0.02	0.20						
High length received	0.01	0.29***	0.29***					
Very-high length received	0.01	0.21***	0.17***	0.21***				
Length provided	0.01	0.50***	0.28***	0.38**	0.49***			
Round	0.01	−0.11***	−0.02	−0.12***	0.09**	−0.11**		
Course	−0.03	−0.47***	−0.41***	−0.47***	−0.50***	−0.60***	0.23***	
Z-score$_J$	0.08*	−0.03	−0.03	−0.05	−0.03	0.08*	0.04	0.05
Low Z-score$_{J-1}$								
Very-low length received	−0.04							
Low length received	−0.08*	0.14***						
High length received	−0.12***	0.15***	0.17***					
Very-high length received	−0.13***	0.25***	0.18***	0.22***				
Length provided	−0.07*	0.51***	0.38***	0.39***	0.51***			
Round	0.04	−0.09**	−0.02	−0.11***	−0.10**	−0.11***		
Course	0.16***	−0.46***	−0.38***	−0.44	−0.51***	−0.65***	0.25***	
Z-score$_J$	0.08*	−0.02	0.002	−0.05	0.06*	0.07*	0.01	0.02
High Z-score$_{J-1}$								
Very-low length received	−0.04							
Low length received	−0.04	0.20***						
High length received	−0.02	0.21***	0.20***					
Very-high length received	−0.02	0.28***	0.18***	0.18***				
Length provided	−0.04	0.48***	0.43***	0.39***	0.47***			

Table 7 (continued)

	Z-score$_{J-1}$	*Very-low length received*	*Low length received*	*High length received*	*Very-high length received*	*Length provided*	*Round*	*Course*
Round	0.06*	−0.12***	−0.04***	−0.11**	−0.10**	−0.14***		
Course	0.09**	−0.49***	−0.39***	−0.44***	−0.51***	−0.65***	0.28***	
Z-score$_J$	0.01	0.02	−0.001	0.001	0.01	0.07*	0.07*	−0.02
Very high Z-score$_{J-1}$								
Very-low length received	0.07*							
Low length received	0.08*	0.22***						
High length received	0.09**	−0.22***	0.15***					
Very-high length received	0.05	0.31***	0.12***	0.22***				
Length provided	0.22***	0.51***	0.45***	0.35***	0.37***			
Round	−0.09**	−0.07*	0.02	−0.09**	−0.07*	−0.07*		
Course	−0.17***	−0.47***	−0.39***	−0.42***	−0.48***	−0.59***	0.22***	
Z-score$_J$	0.15***	0.03	0.01	0.05	0.06*	0.13***	−0.12***	−0.09**

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

Appendix C Stata code for regressions

Regressions in Table 3**:** *bys performance-level: reg Z-Score Z-Score$_{J-1}$ Very-low Length provided$_{J-1}$ Low Length provided$_{J-1}$ High Length provided$_{J-1}$ Very-high Length provided$_{J-1}$ Total Length received$_{J-1}$ i.Course Round, beta*

Regressions in Table 4**:** *bys performance-level: reg Z-Score Z-Score$_{J-1}$ Very-low Length received$_{J-1}$ Low Length received$_{J-1}$ High Length received$_{J-1}$ Very-high Length received$_{J-1}$ Total Length provided$_{J-1}$ i.Course Round, beta*

Data Availability The data presented in this study are available upon request from the corresponding author. The data are not publicly available for reasons of privacy.

Declarations

Conflict of interest disclosure No conflict of interest exists in the submission of this manuscript, and all authors approve the manuscript for publication. We would like to declare on behalf of all co-authors that the work described is original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Ethics approval The study was approved by the University's Human Research Protection Office, and we did not use any data for analysis that contained student names.

References

Abrache, M. A., Bendou, A., & Cherkaoui, C. (2021). Clustering and combinatorial optimization based approach for learner matching in the context of peer assessment. *Journal of Educational Computing Research, 59*(6), 1135–1168.

Adachi, C., Tai, J. H. M., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education, 43*(2), 294–306.

Alemdag, E., & Yildirim, Z. (2022). Effectiveness of online regulation scaffolds on peer feedback provision and uptake: A mixed methods study. *Computers & Education, 188*, 104574.

Ariawan, S. (2018). The effectiveness of cooperative learning method (student team achievement divisions) in Christian education. *International Journal of Education and Curriculum Application, 1*(3), 45–50.

Brown, G. T., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology, 86*(4), 606–629.

Buttaro, A., Jr., & Catsambis, S. (2019). achievement grouping in the early grades: Long-term consequences for educational equity in the United States. *Teachers College Record, 121*(2), 1–50.

Casserly, A. M., Tiernan, B., & Maguire, G. (2019). Primary teachers' perceptions of multi-grade classroom grouping practices to support inclusive education. *European Journal of Special Needs Education, 34*(5), 617–631.

Chang, C. Y., Lee, D. C., Tang, K. Y., & Hwang, G. J. (2021). Effect sizes and research directions of peer assessments: From an integrated perspective of meta-analysis and co-citation network. *Computers & Education, 164*, 104123.

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education, 146*, 103751.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73–84.

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643.

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*(1), 25–36.

Day, I. N. Z., Saab, N., & Admiraal, W. (2022). Online peer feedback on video presentations: Type of feedback and improvement of presentation skills. *Assessment & Evaluation in Higher Education, 47*(2), 183–197.

De Guerrero, M. C., & Villamil, O. S. (2000). Activating the ZPD: Mutual scaffolding in L2 peer revision. *The Modern Language Journal, 84*(1), 51–68.

Francis, B., Taylor, B., & Tereshchenko, A. (2019). *Reassessing Achievement Grouping: Improving Practice for Equity and Attainment*. Routledge.

Giannoukos, I., Lykourentzou, I., Mpardis, G., Nikolopoulos, V., Loumos, V., & Kayafasa, E. (2010). An adaptive mechanism for author-reviewer matching in online peer assessment. In M. Wallace, I. E. Anagnostopoulos, P. Mylonas & M. Bielikova (Eds.), *Semantics in adaptive and personalized services: Methods, tools, and applications* (pp. 109–126). Springer-Verlag.

Gobet, F., & Sala, G. (2023). Cognitive training: A field in search of a phenomenon. *Perspectives on Psychological Science, 18*(1), 125–141.

Hastings, E. M., Krishna Kumaran, S. R., Karahalios, K., & Bailey, B. P. (2022, February). A learner-centered technique for collectively configuring inputs for an algorithmic team formation tool. In L. Merkle & M. Doyle (Eds.), *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1* (pp. 969–975). ACM.

Hattie, J. (2009). *Visible Learning: A Syntheses of over 800 Meta-Analysis Relating to Achievement*. Routledge.

Hattie, J. (2012). *Visible Learning for Teachers: Maximising Impact on Learning*. Routledge.

Henry, T. R. (2013, March). Creating effective student groups: An introduction to groupformation.org. In T. Camp & P. Tymann (Eds.), *Proceeding of the 44th ACM technical symposium on Computer science education* (pp. 645–650). ACM.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111–127.

Hollifield, J. (1987). *Ability grouping in elementary schools*. ERIC Clearinghouse on Elementary and Early Childhood Education.

Hsia, L. H., Huang, I., & Hwang, G. J. (2016). Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. *Computers & Education, 96*, 55–71.

Hsiao, I. H., & Brusilovsky, P. (2011). The role of community feedback in the student example authoring process: An evaluation of annotex. *British Journal of Educational Technology, 42*(3), 482–499.

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' achievement and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101–110.

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' achievement match, feedback quality and essay performance. *Higher Education Research & Development, 36*(7), 1433–1447.

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.

Kanika, Chakraverty, S., Chakraborty, P., & Madan, M. (2022). Effect of different grouping arrangements on students' achievement and experience in collaborative learning environment. *Interactive Learning Environments, 12,* 1–13.

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist, 44*(4), 250–266.

Kim, M. K., & Ketenci, T. (2019). Learner participation profiles in an asynchronous online collaboration context. *The Internet and Higher Education, 41*, 62–76.

Kim, S., Lin, T., Chen, J., Logan, J., Purtell, K. M., & Justice, L. M. (2020). Influence of teachers' grouping strategies on children's peer social experiences in early elementary classrooms. *Frontiers in Psychology, 11*, 587170.

Klein, C., Diazgranados, D., Salas, E., Le, H., Burke, C. S., Lyons, R., et al. (2009). Does team building work? *Small Group Research, 40*(2), 181–222.

Li, H., Bialo, J. A., Xiong, Y., Hunter, C. V., & Guo, X. (2021). The effect of peer assessment on non-cognitive outcomes: A meta-analysis. *Applied Measurement in Education, 34*(3), 179–203.

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.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology, 41*(3), 525–536.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423–458.

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.

Matthews, M. S., Ritchotte, J. A., & McBee, M. T. (2013). Effects of schoolwide cluster grouping and within-class achievement grouping on elementary school students' academic achievement growth. *High Achievement Studies, 24*(2), 81–97.

McKeen, H. (2019). The impact of grade level flexible grouping on math achievement scores. *Georgia Educational Researcher, 16*(1), 48–62.

Min, H. T. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing, 31*, 43–57.

Oppermann, E., Brunner, M., & Anders, Y. (2019). The interplay between preschool teachers' science self-efficacy beliefs, their teaching practices, and girls' and boys' early science motivation. *Learning and Individual Differences, 70*, 86–99.

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science, 43*(5), 591–614.

Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education, 43*(12), 2263−2278.

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education, 27*(3), 534–581.

Sala, G., & Gobet, F. (2020a). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin & Review, 27*(3), 423–434.

Sala, G., & Gobet, F. (2020b). Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Memory & Cognition, 48*(8), 1429–1441.

Saleh, M., Lazonder, A. W., & De Jong, T. (2005). Effects of within-class achievement grouping on social interaction, achievement, and motivation. *Instructional Science, 33*(2), 105–119.

Salehi, N., & Bernstein, M. S. (2018). Hive: Collective design through network rotation. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 151:1–26.

Schmid, U., & Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. *KI-Künstliche Intelligenz, 34*(2), 227–233.

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy, 60*(1), 13–23.

Seifert, T., & Feliks, O. (2019). Online self-assessment and peer-assessment as a tool to enhance student-teachers' assessment skills. *Assessment & Evaluation in Higher Education, 44*(2), 169–185.

Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments, 30*(1), 4–16.

Sheppard, C., Manalo, E., & Henning, M. (2018). Is achievement grouping beneficial or detrimental to Japanese ESP students' English language proficiency development? *English for Specific Purposes, 49*, 39–48.

Shooshtari, Z. G., & Mir, F. (2014). ZPD, tutor; peer scaffolding: Sociocultural theory in writing strategies application. *Procedia-Social and Behavioral Sciences, 98*, 1771–1776.

Slavin, R. E. (1993). achievement grouping in the middle grades: Achievement effects and alternatives. *The Elementary School Journal, 93*(5), 535–552.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of achievement grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research, 86*(4), 849–899.

Storch, N. (2018). Written corrective feedback from sociocultural theoretical perspectives: A research agenda. *Language Teaching, 51*(2), 262–277.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*, 265–269.

Sunahase, T., Baba, Y., & Kashima, H. (2019). Probabilistic modeling of peer correction and peer assessment. In C. F. lynch, A. Merceron, M. Desmarais & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 426–431). ERIC.

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161–1174.
Tsivitanidou, O. E., Constantinou, C. P., Labudde, P., Rönnebeck, S., & Ropohl, M. (2018). Reciprocal peer assessment as a learning tool for secondary school students in modeling-based learning. *European Journal of Psychology of Education, 33*(1), 51–73.
Tsui, A. B. (2011). Teacher education and teacher development. *Handbook of Research in Second Language Teaching and LearniNg, 2*, 21–39.
Urdan, T., & Kaplan, A. (2020). The origins, evolution, and future directions of achievement goal theory. *Contemporary Educational Psychology, 61*, 101862.
Vanhorn, S., Ward, S. M., Weismann, K. M., Crandall, H., Reule, J., & Leonard, R. (2019). Exploring active learning theories, practices, and contexts. *Communication Research Trends, 38*(3), 5–25.
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.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes,* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Eds. and trans.). Harvard University Press.
Wang, D. Y., Lin, S., & Sun, C. T. (2007). Diana: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. *Computers in Human Behavior, 23*(4), 1997–2010.
Wang, Y. Q., & Sun, F. Q. (2018, May). How to choose an appropriate reviewer assignment strategy in peer assessment system? Considering fairness and incentive. In L. Wang, D. Askarany & M. Pawlak (Eds.), *4th Annual International Conference on Management, Economics and Social Development (ICMESD 2018)* (pp. 603–608). Atlantis Press.
Warwick, P., & Maloch, B. (2003). Scaffolding speech and writing in the primary classroom: A consideration of work with literature and science pupil groups in the USA and UK. *Reading, 37*(2), 54–63.
Wen, M., Maki, K., Dow, S., Herbsleb, J. D., & Rose, C. (2017). Supporting virtual team formation through community-wide deliberation. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 533–538.
Wooley, R. S. (2007). The effects of web-based peer review on student writing (Doctoral dissertation, Kent State University).
Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 101826.
Wu, Y., & Schunn, C. D. (2021a). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science, 49*(3), 365–394.
Wu, Y., & Schunn, C. D. (2021b). The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal, 58*(3), 492–526.
Wu, Y., & Schunn, C. D. (2023a). Passive, active, and constructive engagement with peer feedback: A revised model of learning from peer feedback. *Contemporary Educational Psychology, 73*, 102160.
Wu, Y., & Schunn, C. D. (2023b). Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *Journal of Educational Psychology, 115*(1), 118–142.
Xiong, Y., & Schunn, C. D. (2021). Reviewer, essay, and reviewing-process characteristics that predict errors in web-based peer review. *Computers & Education, 166*, 104146.
Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.
Yim, S., & Warschauer, M. (2017). Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology, 21*(1), 146–165.
Yu, S., & Hu, G. (2017). Can higher-proficiency L2 learners benefit from working with lower-proficiency partners in peer feedback? *Teaching in Higher Education, 22*(2), 178–192.
Zhang, X., & McEneaney, J. E. (2020). What is the influence of peer feedback and author response on Chinese University students' English writing performance? *Reading Research Quarterly, 55*(1), 123–146.
Zong, Z., Schunn, C. D., & Wang, Y. (2021). What aspects of online peer feedback robustly predict growth in students' task performance?. *Computers in Human Behavior, 124*, 106924.