Computers in Human Behavior 147 (2023) 107857

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Understanding the what and when of peer feedback benefits for performance and transfer

Qiuchen Yu [a,*], Christian D. Schunn [b]

[a] Faculty of Artificial Intelligence in Education, Central China Normal University, 152 Luoyu Road, Wuhan, 430079, China
[b] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave, Pittsburgh, PA 15260, USA

ARTICLE INFO

Handling Editor: Prof. Nicolae Nistor

Keywords:
Peer feedback
Knowledge transfer
Heterogeneity
Learning
Performance

ABSTRACT

The impact of online peer feedback on learning outcomes has been well-established in previous meta-analyses. However, considerable variations remain unexplained, emphasizing the need to better understand the underlying reasons for heterogeneity. In this study, a large-scale and cross-context dataset, comprising 20,879 students enrolled in 505 assignments within 243 courses at 76 different institutions, was employed. Meta-regression, multi-level modeling, and ANCOVA were used to examine the effect size of each peer feedback experience and explore the associated heterogeneity of effects across courses and assignments within courses. Results revealed that learning benefits are more closely associated with providing feedback rather than receiving it, and are influenced more by the length of peer feedback rather than its helpfulness. Furthermore, heterogeneity exists at the assignment level rather than the course level, and only two peer feedback experiences (i.e., provided length and received length) exhibit particularly large variations in effect size. Particularly novel to our study, we found that learning benefits of both peer feedback experiences diminished as knowledge transfers further away. Similarly, providing feedback demonstrates robust learning compared to receiving feedback, primarily in the form of smaller downward trends and higher learning gains at more further transfer level. Instructors are recommended to design logically-structured consecutive assignments within a course and to provide more guidance to students on giving detailed feedback.

1. Introduction

Peer review is an educational activity in which students respond to the work of their peers (Topping, 1998), which generally consists of quantitative rating process (i.e., peer assessment) and qualitative feedback process (i.e., peer feedback). Over the past 20 years, numerous benefits of peer feedback have been uncovered, such as enhancing performance and learning (K. Cho & MacArthur, 2010; Nelson & Schunn, 2009; Patchan et al., 2016), promoting cognitive and metacognitive skills (Topping, 1998), fostering social skills (Topping, 2009), and improving motivational outcomes (Li et al., 2021). Recent meta-analyses have further confirmed the substantive benefits of peer feedback across a wide range of contexts (Double et al., 2020; Huisman et al., 2019; Li et al., 2016, 2020; Yan et al., 2022), demonstrating that it can yield greater learning benefits than teacher feedback, particularly with regard to academic performance (Double et al., 2020).

Due to these advantages, peer feedback has been increasingly used as an effective pedagogical tool, and notably it is frequently implemented in web-based formats to reduce teachers' workload and better promote student learning through scaffolds afford by the computerized environment (Double et al., 2020; Li et al., 2016, 2020; Sanchez et al., 2017; van Popta et al., 2017; Yan et al., 2022). It generally comprises two main components: receiving feedback and providing feedback, both of which are argued to offer valuable learning opportunities for students (Gielen et al., 2010; Martin & Sippel, 2021; Tsivitanidou et al., 2011). For instance, receiving feedback can encourage students to concentrate on areas requiring improvement and to adopt a reader's perspective, while providing feedback can encourage students to think critically, better understand the rubric and reflect on their work (Nicol et al., 2014). Existing research has examined their separate effects on performance (i.e., do documents improve following peer feedback on the documents?) and learning (i.e., does student performance improve in later assignments?), although more studies have focused on performance (Wu & Schunn, 2021). Notably, when these effects are directly compared in the context of learning outcomes, the benefits of providing versus receiving feedback are sometimes found to be relatively equal (Huisman et al.,

* Corresponding author.
E-mail addresses: yqc@mails.ccnu.edu.cn (Q. Yu), schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.chb.2023.107857
Received 7 May 2023; Received in revised form 18 June 2023; Accepted 20 June 2023
Available online 29 June 2023

2018) and at other times favor providing (Lundstrom & Baker, 2009; Philippakos & MacArthur, 2016). Consequently, there is a pressing need for a large-scale and cross-context dataset to further explore and validate the contribution of peer feedback to student learning as well as to understand how the receiving versus providing aspects of peer feedback experiences shape student learning. Here again, web-based peer feedback is particularly useful because of the availability of large and cross-context datasets.

Thus far, it appears that the relative benefits of peer feedback can vary considerably between different contexts or situations, as consistently evidenced by meta-analyses finding large heterogeneity of effects (Double et al., 2020; Li et al., 2020; Yan et al., 2022). Even though some heterogeneity of effects can be explained by contextual factors (e.g., education level, discipline, and feedback frequency), the meta-analyses still revealed a high amount of unexplained variance within contextual subgroups (Double et al., 2020; Li et al., 2020; Lv et al., 2021). Additionally, recent exploratory work has shed light on two different forms of variation at the course level across 13 courses that were examined (Zong et al., 2021b). Course-level variations in the effect sizes of some predictors of learning appeared to be quantitative in nature, characterized by significantly positive effects that range from weak to strong across courses. In contrast, variations for other predictors were qualitative, where the effect was sometimes significantly positive and sometimes significantly negative across different courses. These large variations in effect sizes are important in terms of theoretical mechanisms and practical instructions, highlighting the need to better understand the extent and underlying reasons for heterogeneity of effects across courses and assignments within courses. The current research seeks to document the heterogeneity in a much larger dataset, clarify whether the heterogeneity is across courses or more specific to assignment-by-assignment variation, and explore several possible causes of the heterogeneity.

2. Theoretical background

2.1. Peer feedback experiences

Peer feedback encompasses two distinct yet interrelated components (i.e., providing feedback vs. receiving feedback), each of which has several potential learning benefits (Gielen et al., 2010; Tsivitanidou et al., 2011). Providing feedback has been recognized as a constructive learning activity (Chi & Wylie, 2014; Wu & Schunn, 2023), affording considerable opportunities for the development of high-level skills (van Popta et al., 2017). When providing qualitative comments, students are more likely to develop evaluative judgments (Boud & Molloy, 2013; Liu & Carless, 2006; Nicol et al., 2014), while also strengthening their ability to detect, diagnose, and resolve different issues (Berggren, 2015; Patchan & Schunn, 2015). Meanwhile, providing feedback also stimulates students to engage in self-reflection on their own work and, in turn, leads to more extensive revision behavior (Dunlap & Grabinger, 2003; Ertmer et al., 2007; Wu & Schunn, 2021; Y. H. Cho & Cho, 2010).

Additionally, it is important to note that providing peer feedback may not always yield favorable results. One potential issue is that students may encounter difficulties when reviewing documents of higher quality than their own (Patchan et al., 2016). In such cases, receiving feedback offers students with a valuable opportunity to engage in active learning as it enables them to take actionable steps towards improving their work. For instance, Butler et al. (2013) considered received peer feedback as particularly useful for giving students understandable explanations, suggestions, and solutions, which may lead to more implementation and revisions. In addition, students can learn how to fill in their knowledge gaps from their peers' feedback (Davey, 2011; Vickerman, 2009). However, the quality of the feedback received can be mixed. Some comments might be detailed, reasonable, and feasible. Conversely, other received feedback can be relatively superficial, not understandable, and (rarely) totally incorrect (Huisman et al., 2018; Patchan et al., 2016; Wu & Schunn, 2020a). Therefore, weaknesses or inaccuracy in received peer feedback may limit its learning benefits (Walker, 2015). Further, in multiple-peer feedback (i.e., students reviewing multiple peers' documents rather than just one), students can receive a large amount of feedback. On the one hand, the multiplicity of this feedback can be more persuasive to students (Gao et al., 2019; Wu & Schunn, 2020b). On the other hand, the sheer volume of the feedback can be overwhelming, confusing, and potentially demotivating to students (Hardavella et al., 2017).

Pragmatically, backward evaluation is another critical component of peer feedback that can easily be included as a scaffold to students in web-based peer feedback system: authors assess the quality of the feedback they received on their work from a reviewer (Luxton-Reilly, 2009; Misiejuk & Wasson, 2021). This accountability component in peer feedback brings several learning benefits to both provider and recipient (Luxton-Reilly, 2009; Misiejuk & Wasson, 2021). From the reviewer's perspective, backward evaluation can encourage students to write higher quality comments (Potter et al., 2017). In addition, asking students to judge perceived comment helpfulness may also help students to make further decisions about the quality of their own documents and comments (Tai et al., 2018). As such, students reflect on the helpfulness of comments they receive, develop their metacognitive skills, and apply those skills to their own work, eventually leading to increased growth in learning performance (Misiejuk & Wasson, 2021). From the author's perspective, backward evaluation also helps students engage in the peer feedback process, prompts them to reflect on received feedbacks, ultimately leading to more revisions or even increased quality of their work (Winstone et al., 2017; Yuan & Kim, 2015). Meanwhile, such backward evaluation also provides another lens through which to examine effects of peer feedback: Does the perceived quality of the provided or received feedback explain variation in learning benefits? A recent study that analyzed data from 13 courses found that comment helpfulness appeared to be a meaningful predictor of student learning, although with small effect sizes for both helpfulness of received and helpfulness of provided comments (Zong et al., 2021b).

Taken together, it remains unclear what aspects of peer feedback experience are robustly associated with changes in students' learning and performance. Although previous studies have examined the separate effects of providing and receiving feedback (K. Cho & MacArthur, 2011; Lundstrom & Baker, 2009; Ion et al., 2019; Martin & Sippel, 2021; Wu & Schunn, 2021), most relied on artificial interventions (K. Cho & MacArthur, 2011; Ion et al., 2019; Lundstrom & Baker, 2009). That is, either the reviewers provided their comments but did not receive any peer feedback during this process or the authors received feedback but did not review other students' documents. However, providing and receiving feedback normally occur simultaneously in natural contexts, and the two experienced in conjunction can moderate the effects of each (Zong et al., 2021b). By contrast, multiple regression approaches can be applied to naturalistic data in most course contexts to statistically tease apart the contributions of each as well as test for interactions (Gao et al., 2023; Wu & Schunn, 2021; Zong et al., 2021b). These research gaps highlight the need for a more comprehensive investigation of predictors of learning in naturalistic data across contexts using a regression approach, rather than relying solely on small number of cases involving non-naturalistic experiments.

2.2. Heterogeneity in peer feedback benefits

As mentioned earlier, previous meta-analyses of experimental studies have generally demonstrated that peer feedback improves student learning outcomes with higher reliability and validity (Double et al., 2020; Huisman et al., 2019; Li et al., 2016, 2020; Lv et al., 2021; Yan et al., 2022). However, these meta-analyses have consistently observed significant heterogeneity of effects (Double et al., 2020; Huisman et al., 2019; Li et al., 2016, 2020; Lv et al., 2021; Yan et al., 2022). Sub-group analysis has been applied within meta-analysis to try to explain the observed variation in effect sizes across studies; however,

it depends upon the availability of relevant moderators in the meta-analysis dataset. Meta-analyses have typically depended upon relatively macro-level context descriptors available in publications, and it may be that more micro-level details matter (e.g., the contents of the comments produced). Indeed, the various tested moderators had limited explanatory power (Double et al., 2020; Li et al., 2020). For instance, contextual factors such as education level, discipline, anonymity, format, and rater training have been examined, but these factors accounted for relatively little variance. That is, high heterogeneity still existed within subgroups even when considering significant moderating factors (Double et al., 2020).

Against this background, a new question arises: does heterogeneity still exists, and at similar levels, when zooming in on the effects of different peer feedback experiences (e.g., the benefits of providing vs. receiving)? It is crucial to uncover *when* students benefit more from a particular kind of peer feedback experience to provide important insights for instructors to obtain, understand, and implement peer feedback more effectively. In the case of existing research, different instructor and research teams implemented the peer feedback process in slightly different ways, using different peer feedback systems, and under various contexts, which all can influence the learning benefits of peer feedback. Therefore, synthesizing and summarizing empirical evidence from many courses and various assignments within courses within one peer feedback system using meta-regression is a crucial step for researchers seeking to gain a more comprehensive understanding of this issue.

One recent study extensively explored the relative contributions of different peer feedback experience across courses. In particular, Zong et al. (2021b) examined the unique contributions of quantity, depth, and quality of received and provided comments in one assignment on students' changes in task performance into the next assignment. Overall, the length of provided feedback was the strongest predictor of improvement in task performance, whereas the helpfulness of provided feedback as well as the length and helpfulness of received feedback were small positive predictors of improvements in task performance. However, the researchers also observed significant unexplained variation in effect sizes associated with length of provided and received feedback across different courses, as well as in amount of provided and received feedback, even though the mean effect size of those predictors was not significantly different from zero. This dataset was too small (only 13 courses) to explore causes of the effect size heterogeneity. It also did not tease apart whether the heterogeneity predominantly existed across courses or whether it predominantly involved heterogeneity across assignment-specific details (and thus only incidentally across course contexts based on assignment differences). Based on these critical gaps, this study aims to first explore factors at the course and assignment level, applying meta-regression techniques to a large-scale, cross-context dataset. Through this approach, the study seeks to provide new insights into meaningful variation in the learning benefits obtained from peer feedback experiences.

2.3. Knowledge transfer in peer feedback

Looking across the broader literature on peer feedback, particularly web-based peer feedback, a majority of studies have focused on performance, meaning improvements in the document being evaluated by peers, rather than learning, meaning improvements that are apparent in future documents, assignments, or assessments (Wu & Schunn, 2021). In other words, researchers have placed the document at the center of their investigations, aiming to explore whether students made meaningful revisions after comprehensive peer feedback activities. Within this focus on performance, both the amounts of received and provided feedback were associated with an increased likelihood of making further revisions (Y. H. Cho & Cho, 2010; Wu & Schunn, 2021). Upon looking at the nature of web-based peer feedback, it was found that the presence of praise and localization (of problems) in comments increased students' likelihood of implementation (Kerman et al., 2022; Patchan et al., 2016). Moreover, the descriptive (i.e., summary) and constructive features (i.e., recommendations) of received feedback are more closely associated with improvement in argumentative essay writing performance (Kerman et al., 2022). However, it is worth noting that relatively few studies (K. Cho & MacArthur, 2011; Lundstrom & Baker, 2009; Philippakos & MacArthur, 2016; Wu & Schunn, 2021, 2023) have directly measured learning (i.e., improvements observed in different assignment), which involves students navigating the challenges of transferring the lessons learned in one document to new assignments, tasks, and situations. The factors that shape document improvements may be different from the factors that shape learning. For example, although it is clear that documents consistently improve when receiving detailed comments from multiple peers, learning has been more consistently observed when students do more than just receive the comments, but rather act on the comments by making revisions, particularly when the comments contained explanations or the revisions involved complex repairs rather that simple typo corrections (Cho & MacArthur, 2011; Wu & Schunn, 2023). In contrast, providing feedback has been more directly related to learning outcomes (Lundstrom & Baker, 2009; Wu & Schunn, 2021). Another way to think of this performance/learning distinction is through the lens of knowledge transfer. For instance, moving from the first draft to the second draft within the same assignment can be considered very near transfer in comparison to moving from one assignment to the next. Further distinctions can be made within transfer levels. For example, transitioning between assignments on different topics but still the exact same writing genre can be regarded as relatively near transfer, in comparison to assignments that substantially change the nature of assignment task.

One meta-analysis on peer feedback has examined the issue of whether different effects of peer feedback are obtained depending upon the level of transfer (Double et al., 2020). Transfer level was coded into no transfer, near transfer, and far transfer based on the similarity between the peer feedback tasks and the academic performance measures. However, we think it is important to problematize the 'no transfer' framing; because this meta-analysis involved learning outcomes that were always new tasks, we would argue that 'no transfer' case actually did involve a small amount of transfer relative to the case of improvements in the document receiving feedback. Interestingly, the meta-analysis showed significantly larger effect sizes for near transfer relative to far transfer. However, there was also large heterogeneity of effects within each transfer level.

One possible explanation for large remaining heterogeneity in the Double et al. (2020) meta-analysis within transfer levels is that their classification approach concentrates solely on the format of the task rather than on broader consideration of elements of learning (i.e., knowledge, skills, and attitudes). More broadly, characteristics of knowledge transfer can be divided into two categories: content and context, respectively (Barnett & Ceci, 2002). Content mainly focuses on what is transferred, such as knowledge, skills, and attitudes. When it comes to peer feedback, instructors generally design assignments following the course syllabus in which assignments logically build up from one to the next. That build relationship may focus on underlying knowledge, involving assignments structured very differently but that use the same underlying knowledge (e.g., two assignments in a unit, one involving analyzing data and the other focused on practical applications). In these cases, the relative performance similarity between consecutive assignments may serve as a measure of the relatively level of knowledge transfer involved from one assignment to the next; this approach to operationalizing transfer is neutral on whether the transfer is based upon similar knowledge, similar skills, or overall attitudes towards the course/assignment. On the other hand, context-based approaches to transfer focus on when and where knowledge is transferred. Writing tasks in peer feedback are typically genre-driven, including argumentative essays, narrative essays, lab reports, reflective journal writing, and summary writing (Lv et al., 2021). As a result, to capture

the context aspect of transfer for peer feedback in writing, it is crucial to accurately identify, understand, and analyze the assignment genre in order to successfully complete it. In such instances, the similarities and analogies between previous and current assignments become particularly relevant. If consecutive assignment genres have nothing in common, the writing skill learned and ability acquired in one specific genre may not be valuable in another genre, potentially resulting in a far knowledge transfer. In sum, whether conceptualized by content or context overlap, there are open questions about how peer feedback experiences are particularly productive for student learning at different transfer levels.

2.4. The present study and research questions

To understand the effects of specific peer feedback experience and how knowledge transfer is associated with meaningful heterogeneity, we leveraged a large-scale, cross-context dataset using the same online peer feedback system to address three research gaps: First, numerous studies have increasingly explored the distinct roles of providing and receiving feedback (e.g., Lundstrom & Baker, 2009; Wu & Schunn, 2021; Wu & Schunn, 2023), but few focused on the specific peer feedback experiences (Zong et al., 2021b). Second, previous meta-analyses have examined the moderating effects of contextual factors on peer feedback (Double et al., 2020; Huisman et al., 2019; Li et al., 2020, 2021; Yan et al., 2022), but they have not delved deeper into the course and assignment levels. Third, prior research has primarily investigated how peer feedback activities function in the context of near transfer (Double et al., 2020), highlighting the need for further exploration of far transfer contexts. The comprehensive research model addressing three research questions is illustrated in Fig. 1.

Research Question 1. What aspects of peer feedback experiences (amount and helpfulness of received and provided comments) substantially predict changes in students' task performance?

Research Question 2. To what extent is there meaningful heterogeneity in the effect sizes across courses and assignments within courses?

Research Question 3. To what extent is heterogeneity in effect size associated with content overlap between consecutive assignments (empirically-derived transfer level and assignment genre change)?

Recently, the Interactive, Constructive, Active, and Passive (ICAP) framework (Chi & Wylie, 2014) has been applied to explain the relative benefits of receiving and providing feedback (Wu & Schunn, 2023). In the context of peer feedback, receiving feedback can be broadly classified as a kind of passive learning task, with students simply being asked to assimilate the information received in the feedback, or an active learning task if they make the revisions suggested by their peers. By contrast, providing feedback involves constructive learning (i.e., more than just active learning), since students must construct suggestions or explanations as part of providing feedback. Building on this theoretical ICAP analysis, providing feedback is likely to result in greater improvements in task performance compared to merely receiving it (RQ1). Indeed, in existing empirical evidence (K. Cho & MacArthur, 2011; Lundstrom & Baker, 2009; Philippakos & MacArthur, 2016; Wu & Schunn, 2021, 2023; Y. H. Cho & Cho, 2010), receiving feedback rarely contributes to learning unless it is accompanied by subsequent revisions (Wu & Schunn, 2023). By contrast, providing feedback consistently leads to learning, as it is directly related to learning and also has a mediated pathway through revisions (Wu & Schunn, 2021, 2023).

Another relevant theoretical framework to consider is social cognitive theory (Bandura, 1989), with a particular focus on self-efficacy as a critical factor that both changes with experience and feedback and also shapes future performance. In the context of web-based multi-peer feedback, students can receive a large amount of critical feedback, and this feedback can lower their self-efficacy, which in turn can result in lower engagement in future assignments and thus lower performance levels. Such self-efficacy effects could explain why the amount of received feedback was sometimes associated with decreases in performance in later assignments (Zong et al., 2021b). This consideration generally supports the general prediction of larger benefits of providing over receiving, as well as predicting the possibility of overall negative effects when other benefits are reduced (e.g., at higher levels of

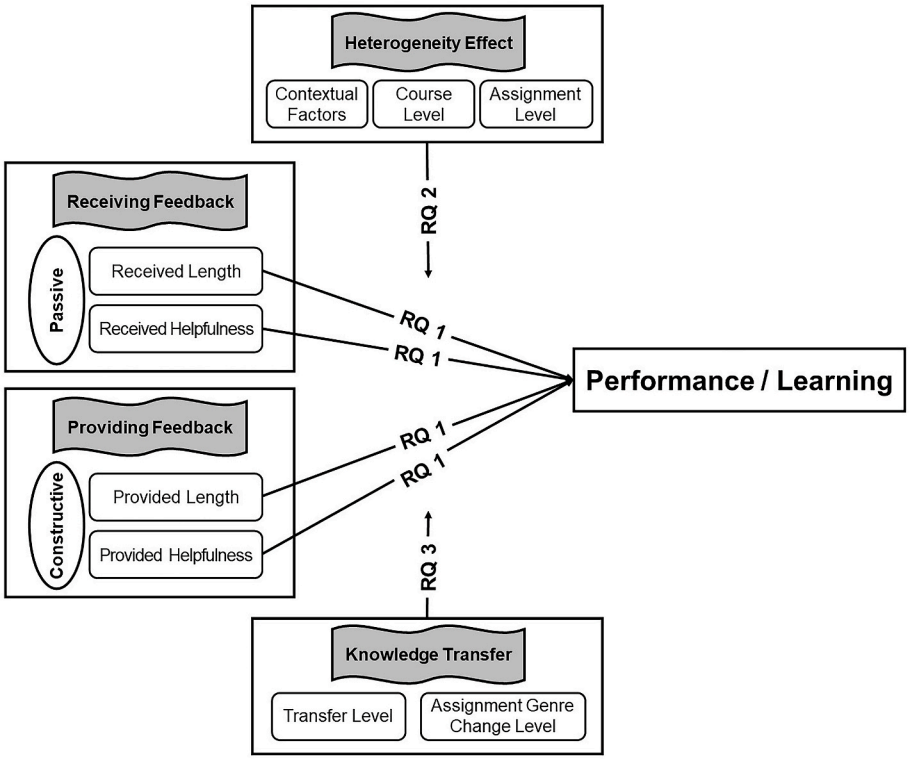

Fig. 1. Research model: Different peer feedback experiences can contribute to the changes in students' performance and learning. Addressing RQ1, receiving feedback is a kind of passive learning, while providing feedback is constructive learning and therefore should have larger effects. Addressing RQ2, heterogeneity is expected in the performance/learning effects of different peer feedback experiences. Addressing RQ3, the benefits obtained from a particular peer feedback experience are likely shaped by knowledge transfer (i.e., transfer level, assignment genre change level).

transfer).

Turning to research question 2, previous meta-analyses of the effects of peer feedback have explored the moderating effects of various contextual factors. However, these contextual factors did not account for much of the heterogeneity (Double et al., 2020; Huisman et al., 2019; Li et al., 2020, 2021). The current study includes a number of contextual factors that have not been previously explored (e.g., course size and # of dimensions), leaving RQ2 as an open research question without any *a priori* predictions about which context factors will be important moderators. However, it is expected that large amounts of heterogeneity will again be observed.

In terms of our knowledge transfer framework, as knowledge transfer becomes increasingly further, there should be lower benefits of both providing and receiving feedback, although it is possible that the declines with increasing levels of transfer are small in the case of providing feedback, since constructive learning may produce more robust learning (RQ3).

3. Methods

3.1. Peer feedback system

Peerceptiv, previously called *SWoRD* (Cho & Schunn, 2007; Schunn, 2016), is an online system that implements multi-peer feedback and has been broadly used worldwide in K-12 and higher education institutions since 2002. Like with almost all multi-peer feedback systems, instructors use *Peerceptiv* to post assignments and corresponding rubrics containing multiple comment prompts and rating dimensions, students upload their own document (e.g., research paper, lab report), and then students review several peers' anonymized documents, providing comments and ratings as directed by the comment prompts and rating rubrics. *Peerceptiv* has also always contained several additional features that improve the quality of ratings and comments, and these features have also been at least partially adopted in many more recently created systems (e.g., *Kritik*, *FeedbackFruits*, *Peergrade*, and *Eli Review*). First, the reliability of the given student's peer ratings is automatically assessed by the system. Second, comment recipients judge the helpfulness of the comments they have received, and these produce an overall helpfulness score for the reviewer (in addition to producing useful data that is also analyzed in the current study). Third, both scores contribute to the reviewer's grade for the assignment. Thus, students are given direct feedback and grading incentives to produce valid ratings and useful comments.

For the time period examined in this study, *Peerceptiv* required instructors to create rating rubrics that used a 7-point Likert scale. A given assignment could have one or more rating rubrics, and typically 3–6 rubrics were included in assignment. The instructor provided brief overall descriptions for each rubric and the option of giving text for each point on the 7-point scale as anchors—we use the text of the rubric description in some of the analyses that examine whether the assignment focus changed from one assignment to the next. In addition to rating dimensions, *Peerceptiv* also has comment dimensions. Each commenting dimension has a brief text description provided by the instructor and then multiple textboxes so that reviewers can provide detailed textual feedback in response to each dimension.

3.2. Ethical considerations

Data were anonymized (e.g., student information is replaced by system numbers rather than names) for analysis by the system prior to researcher access. Use of the anonymized peer feedback data for research was approved by the Human Research Protection Office at University of Pittsburgh. Anonymized educational data of this observational rather than experimental form do not require participant consent.

3.3. Dataset and selection criteria

The dataset was based upon the peer comments and ratings obtained across 505 assignments and produced by 20,879 students enrolled in 243 courses at 76 different institutions, mostly at the university level (see Fig. 2). This dataset was drawn from a larger pool of courses and assignments involving the use of the *SWoRD/Peerceptiv* system between 2010 and 2017; this time period involved a relatively stable user interface and user agreements that enabled data sharing for research purposes. The selected courses and assignments met the following selection criteria. First, the assignment had to be individually submitted (92% of all assignments) rather than a group assignment because it is not possible to track growth in an individual when the task performance represents the work of group. Second, 37% of courses from the database were then selected because 1) they had at least two consecutive individually-completed assignments to support the use of temporally-lagged regression of learning from one assignment to the next, and 2) there were at least 25 students contributing data in those included assignments to a) eliminate test courses (i.e., a fake course in which an instructor or university administrator is evaluating or practicing the functionalities of *Peerceptiv*) and b) also to provide at least moderate power for regression at the level of a single assignment. The resulting courses represented a wide range of disciplines, although science, technology, engineering, and math disciplines (STEM) were the most common (see Fig. 2). Course size (as determined by the number of participating students) also varied widely (i.e., 25–323), although smaller enrollments were most common (see Fig. 2), in part because larger enrollment courses often only had one assignment using peer feedback and thus were excluded. Each course also varied widely in the number of included assignments (i.e., 2–12), although most of the dataset involved the first three assignments since relatively few courses had more assignments. Finally, the assignments also varied widely in terms of how many different comment prompts were included in the reviewing task for a given assignment (i.e., 1–23), although between 3 and 6 included dimensions were most common (see Fig. 2).

We note that the dataset only included information about assignments in the course involving peer feedback. Assignments in the course submitted only to the instructor or for self-assessment could have occurred and would not have been recorded in the database. In addition, other sources of learning such as in-class instruction or discussion, outside-class reading, or guidance from friends or tutoring centers could have occurred. Thus, it was not expected that all learning would be captured by the peer feedback process, as is naturally the case in instruction (i.e., learning typically occurs via multiple pedagogical methods).

3.4. Measures

Analyses were based upon data exported separately for comments data vs. ratings data but in files that aggregated data across all courses. Python scripts were then applied to the raw data files to produce the measures used in the regression analyses (i.e., measures at the level of a given student in a given assignment within a given course). The resulting constructs, variable names, and definitions are summarized in Table 1.

Task Performance. Task performance for each student i on each assignment J within a course, $Z\text{-}Score_{iJ}$, was determined by first calculating the mean value across rating dimensions and reviewers of student i's document, and then standardized within each assignment to $M = 0$ and $SD = 1$ to account for differences in difficulty or rigor of rubrics across assignments.

The use of mean peer ratings as a measure of student task performance deserves some discussion. Expert or instructor ratings could not be obtained in a large-scale study involving tens of thousands of documents. Although not the most common measure of student performance in research, mean ratings from multi-peer assessments do generally have very strong reliability and good validity, and they are now more

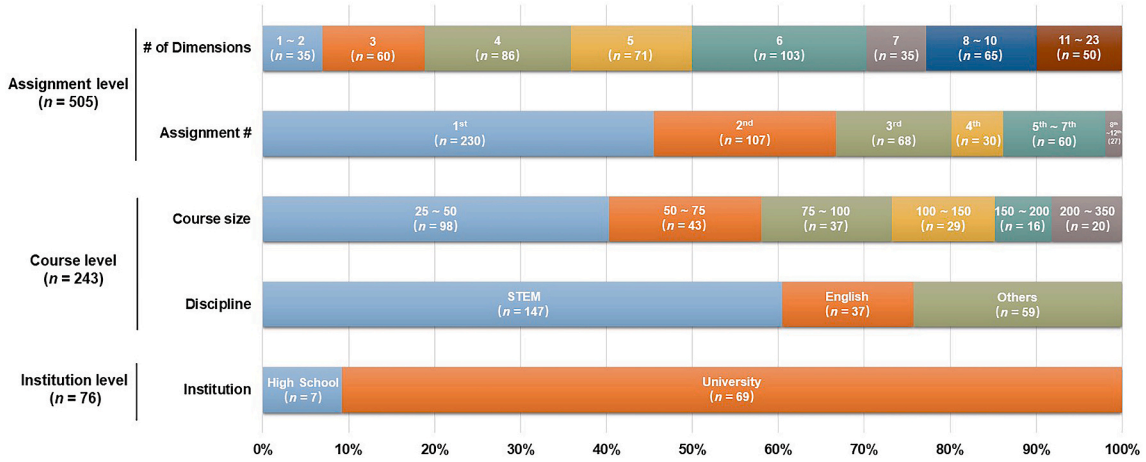

Fig. 2. Frequencies of institution types, course discipline and size, and simple assignment characteristics (which assignment number within the course and how many comment dimensions were included in the peer feedback form).

Table 1
Key student-level constructs, variables names, and definitions.

Construct	Variable	Definition
Task Performance	*Z-Score*$_{iJ}$	The mean score for student *i* on assignment *J* based upon all received peer ratings, standardized to *MD* = 0, *SD* = 1 within each assignment
Provided Comment Length	*Provided Length*$_{iJ}$	The total number of words across comments provided by student *i* on the J^{th} assignment
Provided Comment Helpfulness	*Provided Helpfulness*$_{iJ}$	The mean helpfulness rating across comments provided by student *i* on the J^{th} assignment
Received Comment Length	*Received Length*$_{iJ}$	The total number of words across comments received by student *i* on the J^{th} assignment
Received Comment Helpfulness	*Received Helpfulness*$_{J}$	The mean helpfulness rating across comments received by student *i* on the J^{th} assignment

commonly used in research involving large datasets (Gamage et al., 2021; Suen, 2014). A meta-analysis (Li et al., 2016) found a strong correlation ($r = 0.63$) between mean peer ratings and instructor/expert ratings, and this correlation coefficient was found to be higher when documents were randomly assigned and based upon multiple peers, as in the current context. Further, as noted earlier, *Peerceptiv* implements many best practices as part of its design that tend to produce higher validity ratings (e.g., trait rather than holistic assessment, anchored ratings, double-blind reviewing, requiring comments rather than only ratings, and incentives for higher quality ratings and comments). Studies formally studying the validity of peer assessments in *Peerceptiv* have typically found good reliability and validity of peer ratings across high school, undergraduate, and graduate-level courses, as well as across country contexts (e.g., Cho et al., 2006; Schunn et al., 2016; Zhang et al., 2020). Unpublished analyses of data from the current dataset, focused on data from over a hundred courses in which instructor data was available for at least a subset of student documents, found a median correlation of 0.51 between instructor ratings and mean peer ratings at the level of individual rubrics, which would result in noticeably higher overall task score validity in the typical case of assignments using multiple rubrics (e.g., an estimated overall document score validity of $r = 0.81$ if based upon 4 rubrics with mean individual rubric validity of $r = 0.51$).

Length of Provided/Received Comments. In this study, the length of comments was measured by counting the total number of words used, summed across submitted comments for each comment prompt and across comment prompts within an assignment. Specifically, *Provided Length*$_{iJ}$ was defined as the total number of words involved in the comments provided by a student *i* on the J^{th} assignment. If a student did not complete any reviews, their *Provided Length*$_{iJ}$ was recorded as 0. Similarly, *Received Length*$_{iJ}$ was calculated as the total number of words student *i* received across all reviews on the student's document for the J^{th} assignment. This value was recorded as 0 if a student did not upload a draft for the given assignment, but in practice these cases were excluded from analysis because there was no task score in the absence of a submitted draft.

Sometimes researchers have measured comment length in characters rather than words for several reasons: it is a simpler formula to calculate, to accommodate character-based languages, or to account for increases in idea complexity with longer words (Misiejuk et al., 2021; Zong et al., 2021a). However, character-based analyses have not produced consistent results (Misiejuk et al., 2021; Zong et al., 2021a), the current study only examines comments made in English, and most studies have used the number of words to represent the length of comments (Patchan et al., 2018; Xu & Peng, 2022; Zong et al., 2021b). Therefore, word count rather than character count was used in the measures.

Helpfulness of Provided/Received Comments. In the range of courses being studied, *Peerceptiv* required students to use a specific 5-point Likert scale (1: Not helpful at all – 5: Very helpful) to rate the perceived helpfulness of the comments received on their own documents. Because students sometimes did not complete this required task, there were larger levels of missingness for this variable. *Provided Helpfulness*$_{iJ}$ was defined as the mean helpfulness rating received by student *i* across the comments they provided to their peers' documents on the J^{th} assignment (%missing = 11%), and *Received Helpfulness*$_{iJ}$ was defined as the mean helpfulness of comments received by student *i* on their document for the J^{th} assignment (%missing = 18%). *Provided/Received Helpfulness*$_{iJ}$ was treated as missing if: 1) students did not complete any reviews (i.e., no provided comments and thus no provided helpfulness); 2) did not complete the helpfulness rating task for comments they received for an assignment (i.e., unknown helpfulness of received comments); or 3) did not submit their document in an assignment and thus had no comments to rate (i.e., no received comments and thus no received helpfulness).

Note that data could be systematically missing for an entire assignment if an instructor did not require students to rate helpfulness. In cases of high percentages of missing data at the assignment level (i.e., >50%), we excluded *Provided Helpfulness*$_{J}$ and *Received Helpfulness*$_{J}$ from the regression analysis for that assignment (i.e., the regression coefficients for the other predictors were calculated without controlling for provided and received helpfulness effects).

Content overlap between consecutive assignments. Two measures were created to measure the extent of underlying content overlap between two consecutive assignments. The first measure for content overlap between assignment J and assignment $J + 1$, $Transfer_J$, was a four-level categorical measure based upon the strength of the between-student consistency in relative performance from one assignment to the next. Specifically, for each assignment, we first conducted a linear regression in which task performance on the J^{th} assignment (i.e., $Z\text{-}Score_{iJ}$) alongside experience factors (length and helpfulness of comments provided and received) were predictors of task performance on the $J + 1$st assignment (i.e., $Z\text{-}Score_{iJ+1}$). Then, we categorized $Transfer_J$ into four groups (i.e., very near transfer, near transfer, far transfer, and very far transfer) based upon the effect size of $Z\text{-}Score_{iJ}$ as a predictor. The specific thresholds for the four transfer level groups were defined by the distribution of the estimated effect sizes for $Z\text{-}Score_{iJ}$. Conceptually, this measure takes into account the extent to which similar knowledge, skills, and attitudes are critical for success within two consecutive assignments as measured by the peer assessment rubrics.

From a top-down perspective, the overlap between consecutive assignments was also categorized into four levels in terms of the degree of assignment genre change, D-$Genre_J$, from assignment J to $J + 1$. To identify the change in genre, we first identified the genre for each assignment based upon the short assignment description provided in the dataset. Because some assignment descriptions did not give genre details (e.g., "First assignment"), 205 cases were excluded and 300 cases were codable. Common assignment genres identified in the dataset included argumentative essays, narrative essays, lab reports, reflective journal writing, and summary writing, similar to the common genres reported as being used with peer-evaluation (Lv et al., 2021). Then the levels of assignment change from one assignment to the next was coded based upon these genre details. Table 2 presents detailed descriptions for each of the four levels and examples for each level.

To explore the relationship between assignment genre change and transfer, the relative frequency of transfer level among four different types of assignment genre change is shown in Fig. 3, which illustrates what percentage of assignments at different assignment genre change levels were categorized very near transfer, near transfer, far transfer, or very far transfer. Although the two constructs are clearly related, particularly with more near transfer for same assignment, different draft cases, the two variables turned out to be only loosely related constructs, with a mix of near transfer and far transfer cases within each genre-change level, an interesting outcome that we take up in the general discussion.

Contextual factors. To examine whether observed variations in effect sizes across assignments and courses was related to contextual factors, several factors available in the dataset were examined. One contextual factor was related to a system setup parameter: reviewing style (assigned one-at-a-time vs. assigned all-at-once). The one-at-a-time assignment method entailed providing reviewers with a single document at a time to review, and only providing another document to review after the prior review was completed. This approach allowed for easy integration of late-submitted documents, better balance in the number of completed reviews for each document, and also minimized the number of reviews

Table 2
Levels of assignment genre change, along with an example for each level of change.

Level of change	Example
Same assignment different draft	J^{th}: PHYS Formal Lab Report Draft 1 $J + 1$st: PHYS Formal Lab Report Draft 2
Same genre different topic	J^{th}: Case Study: Religion $J + 1$st: Case Study: Race/Ethnicity
Related genre	J^{th}: Introduction $J + 1$st: Draft Materials & Methods
Unrelated genre	J^{th}: Molecules and Light Investigation Report $J + 1$st: Energy Project

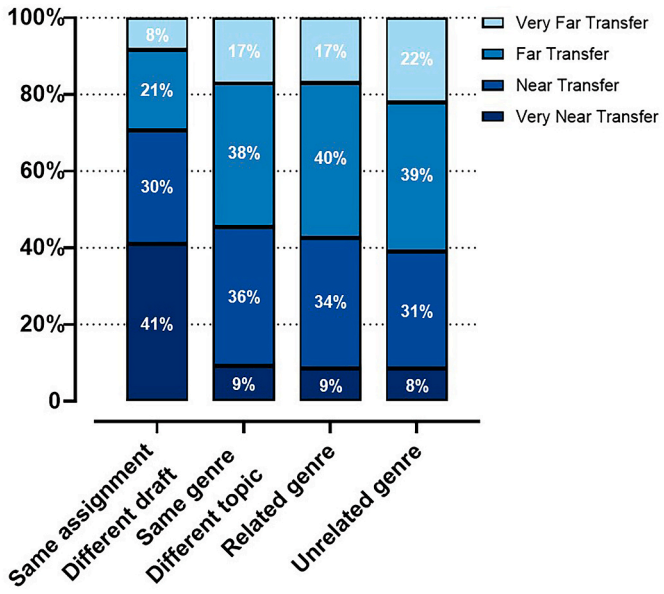

Fig. 3. Relative frequency of transfer level among four levels of assignment genre change.

that were assigned but never completed. In the all-at-once assignment approach, all required documents to review were allocated to reviewers at the beginning of the reviewing phase. This method required assigning late submissions as additional workload for reviewers as well as resulting in a widely varying number of completed reviews for each document. Some contextual factors related to the course context. Institution type (high school vs. university) and discipline (STEM vs. English vs. Others) were hand-coded (as needed) from the institution name and course name in the dataset. Furthermore, course sizes were divided into small (25–50 students), medium (50–100 students), and large (>100 students). Some contextual factors were related to the assignment context. Assignment numbers were classified into three groups based upon relative frequencies: 1st, 2nd–3rd, and 4th–12th. Similarly, rating rubrics were stratified into three categories of roughly similar size (1–3, 4–6, and 7–23) based upon on the number of dimensions in the peer feedback form.

Assignment-level distributional properties. Because statistical properties of dependent and independent variables within each assignment can influence regression models, the mean, standard deviation, skewness, and kurtosis within each assignment were calculated for the dependent variable and all four predictors (i.e., length and helpfulness of comments received and provided) to examine whether the statistical properties of the variables biased the estimated effect sizes (e.g., were some effect sizes lower simply because the number of submitted or received comments in that assignment were much smaller on average or varied less across students?).

3.5. Data analysis procedure

Calculating assignment-specific regression coefficient. To ultimately establish a dataset for meta-regression analysis, as a first step, separate time-lagged multiple linear regressions of changes in relative task performance were conducted within partial datasets, each unique to a specific assignment within a specific course. These regression datasets involved the student-level measures derived from the raw rubric ratings, comments, and helpfulness ratings. Specifically, for each assignment except the last assignment in a course (i.e., 505 cases), we conducted an assignment-specific multiple linear regression with $Provided\ Length_{iJ}$, $Provided\ Helpfulness_{iJ}$, $Received\ Length_{iJ}$, and $Received\ Helpfulness_{iJ}$ used as key independent variables, $Z\text{-}Score_{iJ}$ as the base-line variable, and $Z\text{-}Score_{iJ+1}$ as the dependent variable (i.e., 505 different regressions). This

process produced a β coefficient and standard error for *Z-Score*, *Provided Length*, *Provided Helpfulness*, *Received Length*, and *Received Helpfulness* for every assignment (i.e., approximately 5,000 values were estimated—5 x 2 x 500). Note that due to the high missingness of helpfulness ratings in a few assignments, 44 linear regression models did not include *Provided Helpfulness*$_{iJ}$ and *Received Helpfulness*$_{iJ}$ as predictors and thus their corresponding β and standard errors were not estimated in those 44 assignments. Since the smaller enrollment courses, by definition, involved somewhat low *N*s for multiple regression involving 5 predictors, even modest collinearity may have caused problems. However, examination of Variance Inflation Factors revealed that significant multicollinearity problems were not observed in any of the regressions (i.e., all *VIF*s < 3).

To illustrate the process for producing β coefficients, we selected four cases, each based upon one assignment in one course. The selected cases also illustrate the kinds of variation that was often observed in the relationships between *Provided/Received Length*$_{iJ}$ and *Z-Score*$_{iJ+1}$, (in the multiple regressions that control for other peer feedback experience predictors). The values in each scatterplot were produced using the "margins" command in Stata 17, which estimated marginal means for a particular predictor (i.e., *Z-Scores*$_{iJ+1}$) at every unique value of the predictor. Note that in the rare cases in which there were multiple student cases at an exact given x-axis value, the mean y value is displayed. The four cases depicted in Fig. 4 respectively illustrate a strong positive β, a significant but weaker positive β, a null β, and a significant negative β, using data drawn from four larger courses. The top two show typical patterns for provided length across assignments and the bottom two show typical patterns for received length, although the weak positive case also occurred regularly for received length.

Meta-regression. To address the first research question, we conducted random-effects meta-regressions with restricted maximum likelihood using the "meta" command in Stata 17 (Langan et al., 2019), separately for each predictor. The primary outcome was the mean effect sizes of each independent variable (i.e., *Provided Length*$_{iJ}$, *Provided Helpfulness*$_{iJ}$, *Received Length*$_{iJ}$, and *Received Helpfulness*$_{iJ}$) in the multiple linear regression models (e.g., what is the average effect size of the relationship of *Provided Length*$_{iJ}$ with growth in task performance, *Z-score*$_{iJ}$, from one assignment to the next?). Subsequently, sub-group analysis was employed to identify possible contextual factors (i.e., reviewing type, institution type, discipline, course size, assignment#, and # of dimensions) that may influence each of the effect sizes associated with each of the four predictor variables (e.g., is the effect size for a given predictor larger in university courses than high school courses?).

To further investigate whether variation in effect sizes across assignments/courses was due to poorly estimated coefficients (e.g., from smaller sample sizes or noisy estimates of task performance), we examined the meta-analysis I^2 statistic, which indicates the percentage of variation in effect sizes across assignments that is not attributable to noise (i.e., as indicated by relatively large standard errors). Specifically, I^2 values of 0%–25%, 25%–50%, 50%–75%, and 75%–100% are considered unimportant, low, moderate, and high heterogeneity, respectively (Higgins et al., 2003). Moreover, we also conducted a linear correlation analysis to test whether the distributional properties of the predictor or outcome variables within an assignment were significantly correlated with each effect size.

Multi-level modeling. Furthermore, to address the second research question, a null multi-level model (level 1: assignment, level 2: course) was tested for each predictor's effect size estimate in each assignment, and the intraclass correlation coefficient (ICC) was examined to establish the relative amount of variance at the assignment vs. course levels (i.e., did effect sizes vary more across different courses or more across assignments within a course?). The results of this analysis also influenced the choice of analytic approach to address the third research question: methods that adjusting for nesting effects, like multi-level modeling, need to be considered only when the level 2 ICC is greater than 10% (Lee, 2000).

Test of transfer level and genre-change effects. Finally, to address research question 3, we conducted a series of ANCOVAs to examine how content overlap directly contributes to these heterogeneity effects. Specifically, *Transfer*$_J$ and D-*Genre*$_J$ were used as categorical variables (in separate ANCOVAs), and the effect sizes of different peer feedback experiences were included as dependent variables. Nested models of assignments within courses were not used due to the lack of significant

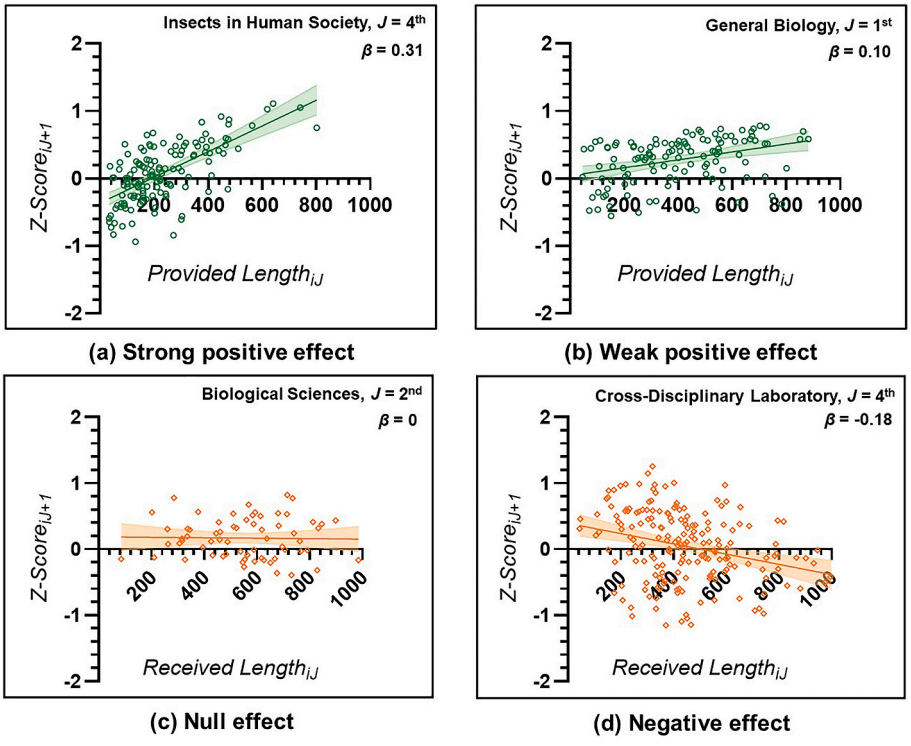

Fig. 4. Four example specific assignment/course cases of the common patterns observed in the 505 multiple linear regressions.

course-level variance. Additional contextual factors were included as categorical variables in the ANCOVAs in the case of significant contextual variation in effect sizes (e.g., reviewing style, course size, assignment#, and # of dimensions). Similarly, assignment-level statistical distributional properties (e.g., variable kurtosis) were included as covariates in the ANCOVAs when significant correlations of the distributional properties of a given predictor were identified with that variable's corresponding effect size.

4. Results

As a first step towards validating the appropriateness of the assignment-specific multiple regression approach, Table 3 presents means, standard deviations, and meta-correlations at the student level of each the peer feedback experience predictors (e.g., what is the average correlation between a student's Z-score and their provided length within an assignment?). The student-level variables were generally correlated with one another at relatively weak levels, which is consistent with the relatively low VIF values for all assignment-level multiple-regressions. However, the correlations were not zero. Thus, the multiple-regression approach was both necessary and appropriate.

A secondary sanity check involved the independence of the variable-specific effect sizes. It was possible that all effect sizes might be larger in some courses and smaller in other courses due to lack of any measurable learning effects in some courses or high noise in the task performance measures. Additionally, correlations among these effect size values for each peer experience predictor (i.e., standardized betas in multiple linear regression) are shown in Table 4, alongside mean and standard deviations in effect sizes. The *ES Z-Score$_{iJ}$* correlated significantly but at modest levels with all of other peer feedback experience effect sizes, except with *Provided Length$_{iJ}$*. Thus, it is unlikely that variation in effect sizes for each experience predictor is based upon poor task measurement. Furthermore, *ES Provided Length$_{iJ}$* demonstrated a modest but significant positive correlation with *ES Received Helpfulness$_{iJ}$*. All other effect sizes were not significantly correlated with one another. In general, the modest intercorrelations among predictors suggests they have separable factors underlying their variation (if they are established to be meaningful variation).

4.1. What aspects of peer feedback experiences substantially predict changes in students' task performance?

The key statistics from the meta-regressions for each regression effect size are presented in Table 5. Unsurprisingly, task performance in one assignment had the strongest association with task performance in the next assignment. However, it was important to establish that, in these assignments, it was generally useful to include student performance on the previous task as a baseline. This information was also one step in establishing that the transfer measure was meaningful in these assignments. Among the peer feedback experience variables, the two predictors related to providing feedback were significantly different from zero (and positive). Further, provided length in particularly was a robust predictor of improvements in students' task performance, with a mean effect size more than three times greater than the effect size associated with provided comment helpfulness. These results provided additional evidence for the dominant role in general of providing feedback over receiving it in task learning.

Statistically significant and large heterogeneity in effect sizes was found for three of the five effect sizes: *Z-Score$_{iJ}$*, *Provided Length$_{iJ}$*, and *Received Length$_{iJ}$*. Contextual factors were tested as possible moderators for those three effect sizes showing significant heterogeneity (see Appendix Tables A1-A3 for complete details). On the whole, there was some significant contextual moderation, but the moderation was generally small and the general overall patterns for each effect sizes held within each context: large positive mean values for *ES Z-Score$_{iJ}$*, smaller positive mean values for *ES Provided Length$_{iJ}$*, and near zero mean values for *ES Received Length$_{iJ}$*.

Regarding the overall effect of *Z-Score$_{iJ}$* (see Table A1), institution type ($p = .37$), discipline ($p = .46$), and assignment # ($p = .10$) did not exhibit significant contextual moderation. However, course size ($p = .012$), # of dimensions ($p = .036$), and reviewing type ($p < .001$) were found to be statistically significant moderators: *ES Z-Score$_{iJ}$* was larger when peer feedback was implemented in smaller course sizes, when instructors adopt a more multidimensional rating rubric, and when documents were assigned all-at-once. At the same time, these moderation effects were generally small and there continued to be statistically significant and large heterogeneity for these effect sizes within every context. These findings for *ES Z-Score$_{iJ}$* are consistent with treating variation in the effect size for Z-score as a measure of the relative amount of transfer from one assignment to the next rather than a simple contextual confound.

Turning to the effect size of *Provided Length$_{iJ}$* (see Table A2 for details), only # of dimensions ($p = .007$) was a statistically significant moderator, with a slightly larger effect size in the case of a medium number of rating dimensions. The other contextual moderators were not statistically significant—institution type ($p = .90$), discipline ($p = .45$), course size ($p = .31$), assignment# ($p = .07$), or reviewing type ($p = .052$). Further, there was statistically significant and large heterogeneity of effect size within every context.

Finally, none of the contextual factors were statistically significant moderators of *ES Received Length$_{iJ}$*—institution type ($p = .15$), discipline ($p = .70$), course size ($p = .91$), assignment# ($p = .38$), # of dimensions ($p = .67$), and reviewing type ($p = .46$). Further, there was statistically significant and large heterogeneity in effect size in every context. Therefore, for this effect size, as with the other experience predictor showing heterogeneity, there was an opportunity for other factors like genre change or different levels of underlying cross-assignment transfer to explain variation.

Table 3
Variable means and standard deviations, and mean estimated linear correlations among variable at the student level within an assignment. Mean correlation values are based upon meta-correlation analyses across n = 505 assignments.

	Mean	SD	2	3	4	5
1. *Z-Score$_{iJ}$*	0	1	0.21***	0.10***	−0.15***	0.15***
2. *Provided Length$_{iJ}$*	501	653		0.18***	−0.04***	−0.06***
3. *Provided Helpfulness$_{iJ}$*	3.87	0.95			−0.01	−0.01
4. *Received Length$_{iJ}$*	501	496				0.06***
5. *Received Helpfulness$_{iJ}$*	4.28	0.70				

Notes. *$p < .05$; **$p < .01$; ***$p < .001$.

Table 4
Mean, standard deviation, and linear intercorrelations among the assignment-level effect sizes for each of the predictors.

	Mean	SD	2	3	4	5
1. *ES Z-Score$_{iJ}$*	0.26	0.23	−0.08	−0.13**	0.17***	−0.17***
2. *ES Provided Length$_{iJ}$*	0.11	0.16		−0.08	0.01	0.16***
3. *ES Provided Helpfulness$_{iJ}$*	0.04	0.17			−0.02	0.01
4. *ES Received Length$_{iJ}$*	−0.01	0.18				0.05
5. *ES Received Helpfulness$_{iJ}$*	−0.001	0.18				

Notes. *ES* = effect size. *$p < .05$; **$p < .01$; ***$p < .001$. *Ns* vary between 461 and 505.

Table 5
Meta-analysis results for the overall effect of each peer feedback experience, as well as the heterogeneity of corresponding effects (significant overall effects and heterogeneity of effects in bold).

Predictor	*Overall Effect*				*Heterogeneity of Effect*		
	Effect Size	*95% CI*	*Z*	*p*	χ^2	*p*	I^2
Z-Score$_{iJ}$	**0.27**	[0.25, 0.29]	28.2	**< .0001**	**1307.9**	**< .0001**	**62%**
Provided Length$_{iJ}$	**0.11**	[0.10, 0.13]	16.1	**< .0001**	**1.60 x 10^8**	**< .0001**	**100%**
Provided Helpfulness$_{iJ}$	**0.04**	[0.03, 0.06]	6.4	**< .0001**	383.4	0.99	2%
Received Length$_{iJ}$	−0.01	[-0.02, 0.01]	−0.7	0.46	**7.00 x 10^7**	**< .0001**	**100%**
Received Helpfulness$_{iJ}$	−0.001	[-0.017, 0.015]	−0.1	0.9	276.7	1	0%

4.2. Non-content sources of heterogeneity in effect sizes

As noted in the prior section, there was large and pervasive heterogeneity for three of the effect sizes. The significant heterogeneity in the effect size of *Z-Score$_{iJ}$* forms the basis for the transfer measure, while the significant heterogeneity in the effect sizes for *Provided Length$_{iJ}$* and *Received Length$_{iJ}$* form the basis of testing the impact of relative transfer and genre change on those variables. To illustrate these large variations in effects, Fig. 5 depicts the distribution of these estimated effect sizes. The variation in *ES Z-Score$_{iJ}$,* and *ES Provided Length$_{iJ}$* is predominantly quantitative: significant and positive in most cases, but of varying size. Conversely, the variation in *ES Received Length$_{iJ}$* is more qualitative: the effects were sometimes positive and sometimes negative, but also many near-zero/not statistically significant values.

To investigate the relative amount of heterogeneity at the course level, we employed multilevel null models of assignments nested within courses and calculated the intra-class correlation coefficients (ICCs) at the course level. 19% of the variance in *ES Z-Score$_{iJ}$* was associated with course-level differences, leaving 81% of the variability at the assignment level. Conversely, the estimated ICCs for the *ES Provided Length$_{iJ}$*, *ES Provided Helpfulness$_{iJ}$*, *ES Received Length$_{iJ}$*, and *ES Received Helpfulness$_{iJ}$* were all below 0.01. These findings indicated minimal variability was due to courses factors, and instead observed effect size variation was entirely at the assignment level.

One possible source of variation in effect sizes at the assignment level was essentially statistical artifacts from statistical properties of predictor or outcome variables (e.g., restricted range effects or violations of normality). *ES Z-Score$_{iJ}$* and *ES Provided Length$_{iJ}$* were significantly associated with kurtosis values, but these coefficients were relatively small (i.e., absolute values all less than 0.12, linear correlations $\leq$10% of variance; see Table A4 for details). No other associations were significant. Therefore, the large heterogeneity in peer feedback experiences' effect sizes were unlikely to be explained by statistical artifacts.

4.3. To what extent is heterogeneity in effect size associated with degree of content overlap?

Two content-based approaches were taken to explain variation in effect sizes: transfer levels and genre change. The results of the ANCOVA analyses examining effect size variation as a function of transfer level are shown in Table 6. The first row, *ES Z-Score$_{iJ}$*, is by definition different across transfer levels and is simply presented to provide information about relative size of group differences. In the analysis of *ES Z-Score$_{iJ}$*, course size, # of dimensions, and reviewing type were included as categorical variables, while *Kurtosis of Z-Score$_{iJ}$* was included as a covariate control.

Table 6
Mean effect sizes and standard deviations within each transfer level. Significant effects in bold.

Variables	Very near transfer (n = 84)	Near transfer (n = 150)	Far transfer (n = 179)	Very far transfer (n = 92)
ES Z-Score$_{iJ}$	0.59 ± 0.09^a	0.37 ± 0.06^b	0.17 ± 0.06^c	−0.07 ± 0.13^d
ES Provided Length$_{iJ}$	0.08 ± 0.15	0.12 ± 0.16	0.12 ± 0.14	0.13 ± 0.19
ES Provided Helpfulness$_{iJ}$	0.01 ± 0.18	0.03 ± 0.16	0.03 ± 0.15	0.08 ± 0.20
ES Received Length$_{iJ}$	0.03 ± 0.22^a	0.003 ± 0.15^a	−0.005 ± 0.15	−0.06 ± 0.21^b
ES Received Helpfulness$_{iJ}$	−0.03 ± 0.17	−0.01 ± 0.15	−0.01 ± 0.15	0.05 ± 0.26

Note. Different letters within a row indicate significant differences among groups at α = 0.05 using Tukey post-hoc tests.

More relevant to research question 3, the transfer groups also differed significantly in terms of *ES Received Length$_{iJ}$*, $F_{(3, 501)} = 4.35$, $p = .005$. Generally speaking, the further the transfer level, the more negative the effect size of *Received Length$_{iJ}$*, and positive values of *Received Length$_{iJ}$* were largely reserved for the nearest transfer cases. Subsequent post-hoc Tukey tests demonstrated that the very near transfer group and near transfer group had a higher *ES Received Length$_{iJ}$* than the very far transfer group (see Fig. 6).

The second ANCOVA analysis focused on the relationship of effect sizes with assignment genre change (see Table 7). The first row simply replicates what was shown in Fig. 3: much higher *ES Z-Score$_{iJ}$* values in the next draft. Of the four experience predictors, only *ES Provided Length$_{iJ}$*, showed a statistically significant relationship with genre change, $F_{(3, 293)} = 4.83$, $p = .002$ (see Fig. 7). The ANCOVA analysis of this variable included # of dimensions as a categorical variable and *Kurtosis of Provided Length$_{iJ}$* as a covariate control, although the genre-change results are unaffected by their inclusion. More specifically, *ES Provided Length$_{iJ}$* was small in different draft cases but moderate in same genre but different topic cases.

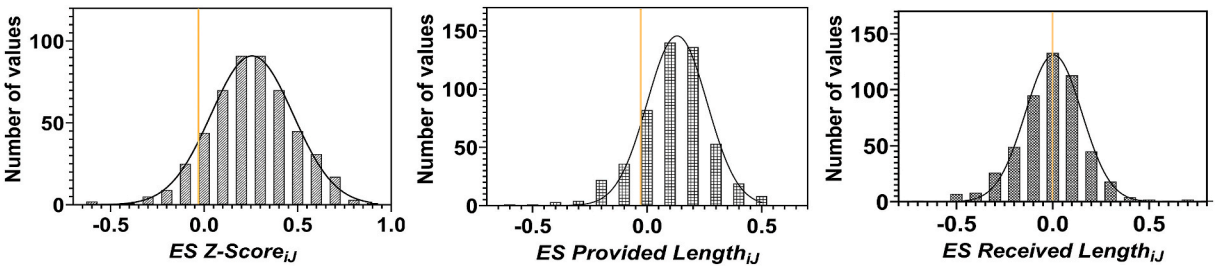

Fig. 5. Distribution of the estimated effect sizes for *Z-Score$_{iJ}$, Provided Length$_{iJ}$*, and *Received Length$_{iJ}$*.

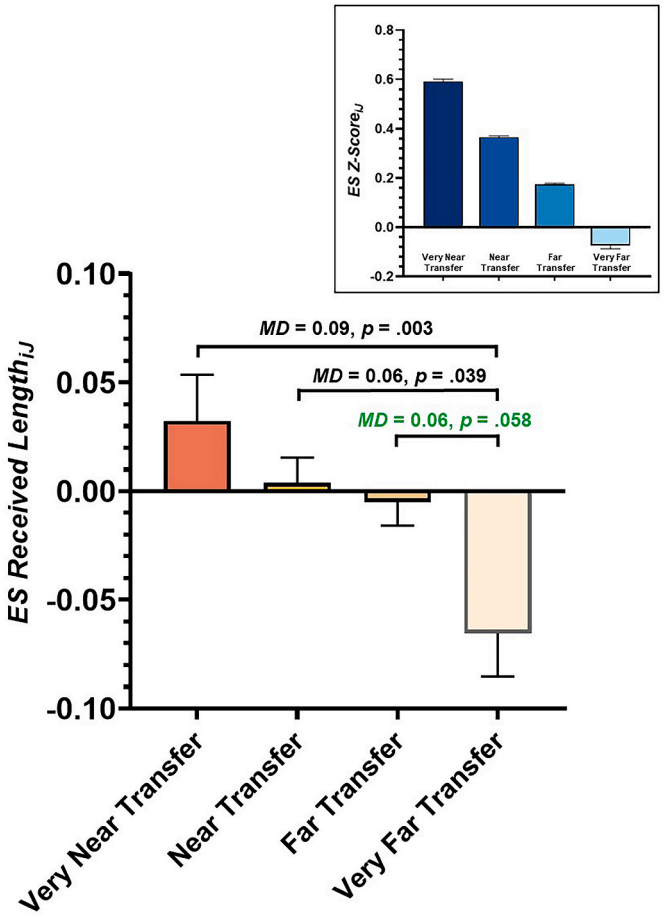

Fig. 6. Means (with SE bars) of *ES Received Length*$_{iJ}$ at each of the four different transfer level; inset show the basis of the transfer levels. Significant group differences are indicated based upon Tukey post-hoc correcting for multiple comparisons.

Table 7
Mean effect sizes and standard deviations within each level of assignment genre change. Significant effects in bold.

Variables	Same assignment different draft (n = 95)	Same genre different topic (n = 99)	Related genre (n = 47)	Unrelated genre (n = 59)
***ES Z-Score*$_{iJ}$**	0.39 ± 0.24[a]	0.25 ± 0.20[b]	0.22 ± 0.23[b]	0.22 ± 0.20[b]
***ES Provided Length*$_{iJ}$**	0.06 ± 0.18[b]	0.16 ± 0.15[a]	0.11 ± 0.16	0.11 ± 0.13
ES Provided Helpfulness$_{iJ}$	0.05 ± 0.20	0.06 ± 0.13	−0.01 ± 0.21	0.02 ± 0.13
ES Received Length$_{iJ}$	−0.03 ± 0.20	0.02 ± 0.14	−0.01 ± 0.20	−0.02 ± 0.17
ES Received Helpfulness$_{iJ}$	0.01 ± 0.15	0 ± 0.16	0.03 ± 0.27	−0.01 ± 0.16

Note. Different letters within a same row indicate significant differences among groups at α = 0.05 using Tukey post hoc tests.

5. Discussion

Using authentic web-based peer feedback data from a large sample of assignments (N = 505) collected from a specific peer feedback system, we assessed the effect-sizes of the relationships of each peer feedback experience with performance and learning. In addition to drawing general conclusions, we examined whether the heterogeneity stems

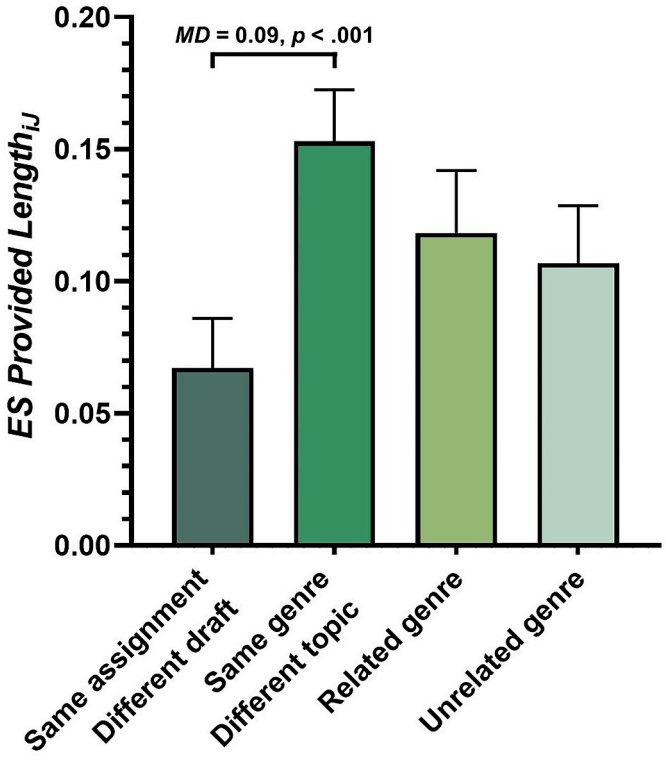

Fig. 7. Marginal means (with SE bars) of *ES Provided Length*$_{iJ}$ at each of the four levels of assignment genre change. Significant group differences are indicated based upon Tukey post-hoc correcting for multiple comparisons.

from the course or the assignment level as well as exploring the relationship between knowledge transfer and meaningful heterogeneity. Our main findings are summarized in Fig. 8. Overall, providing feedback, rather than receiving it, was more closely associated with learning outcomes, and relationship with length was greater than the relationship with helpfulness in the providing aspect. As expected, a considerable amount of heterogeneity existed in the effect sizes of these relationships, and these analyses revealed that this effect-size heterogeneity is actually at the assignment level rather than the course level. Further, some of the heterogeneity (i.e., for provided length and received length effects) could be partially explained by relative levels of knowledge transfer. Thus, the present study contributes to the existing research by showing generality of findings across a large data set (i.e., confirming the differential effects of specific peer feedback experiences), clarifying the nature of heterogeneity of effects, and providing new explanations of the heterogeneity observed in previous meta-analyses.

5.1. RQ1. What aspects of peer feedback experiences substantially predict changes in students' task performance?

Comprehensive examinations of individual peer feedback experiences, utilizing a large naturalistic dataset of web-based peer feedback, confirmed our proposed hypotheses. As shown in Table 5, the meta-analysis results provided straightforward evidence that learning benefits from peer feedback are much more strongly associated with the providing aspect rather than the receiving aspect of peer feedback. Moreover, the quantity of provided feedback (i.e., provided length) was found to substantially predict improvements in students' task performance, whereas the quality of provided feedback (i.e., provided helpfulness) played a lesser role. The present study confirmed the generality of these patterns that were previously obtained from peer feedback tasks carefully constructed by researchers (Wu & Schunn, 2021, 2023; Zong et al., 2021b).

Overall, the findings for the first research question were consistent

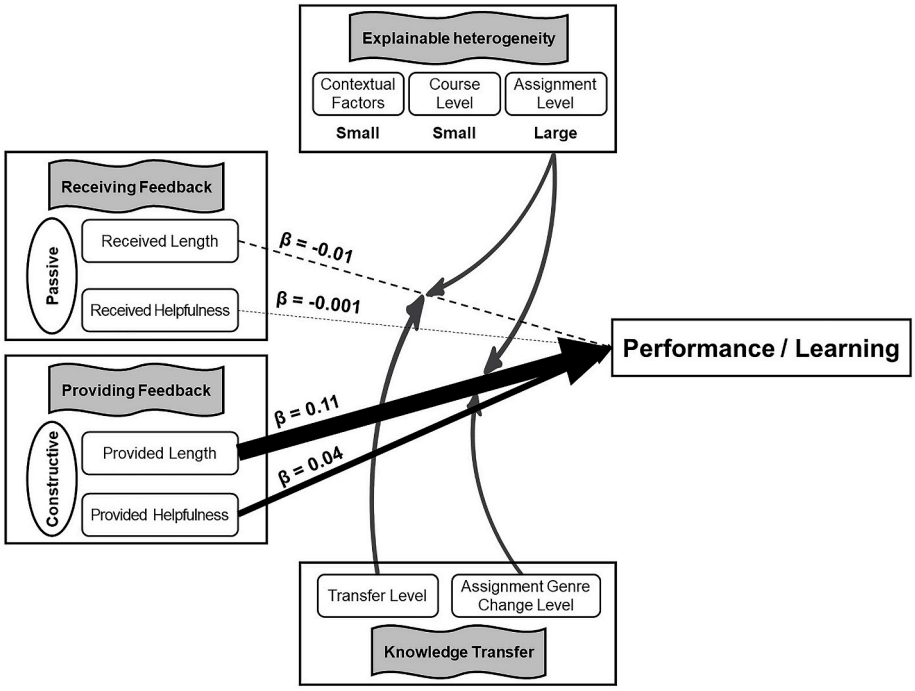

Fig. 8. Revised research model based on main findings. Line thickness corresponds with statistical strength of relationship in regression models, and dotted line indicates a negative effect. The curved arrows at the top indicate cases of significant heterogeneity within the peer feedback experiences, and the curved arrows at the bottom ones indicate which measure of knowledge transfer moderated their benefits.

with the proposed theoretical perspectives of ICAP and social cognitive theories. As expected, receiving feedback, categorized as passive learning, showed weaker benefits than providing feedback, categorized as constructive learning. Further, also consistent with ICAP, it is the nature of the learning work (i.e., the amount of different types of reviewing work) rather than the quality of the learning work (i.e., the helpfulness of the review work) that is associated most strongly with learning outcomes. In addition, consistent with social cognitive theories, receiving large amounts of negative feedback can harm future performance, presumably through reductions in self-efficacy. The combination of the two theoretical accounts are needed to explain why the overall effects are sometimes positive and sometimes zero or negative (i.e., a combination of learning and de-motivation effects).

5.2. RQ2. To what extent is there meaningful heterogeneity in the effect sizes across courses and assignments within courses?

The observed large heterogeneity in effect sizes of peer feedback experiences is consistent with the findings of previous meta-analyses (Double et al., 2020; Huisman et al., 2019; Li et al., 2016, 2020; Yan et al., 2022). Going beyond prior research and consistent with predictions of the ICAP framework, the heterogeneity was found to occur predominantly at the assignment level rather than at the course level, reflecting the cognitive nature of what students are being asked to do. Nevertheless, it's also worthwhile to delve deeper into the heterogeneity at the context level and specifically consider student characteristics. Prior research has shown that feedback performance can be influenced by individual characteristics such as gender (Noroozi et al., 2018, 2022), epistemic beliefs (Noroozi, 2022), feedback literacy (Carless & Boud, 2018; Yan & Carless, 2022), and attitudes towards feedback (Dong et al., 2023; Kasch et al., 2022).

Although not the direct focus of the current study, the heterogeneity in the effect size of prior task performance requires additional discussion. The sub-group analysis found that some contextual factors (i.e., course size, # of dimensions, reviewing type) could explain this heterogeneity, but their explanatory power diminished when controlling for assignment genre changes. This observation suggests that significant moderators might be confounded due to the presence of more typical assignment genre changes in some subgroups, such as different drafts of the same assignment or different topics for the same genre.

The current study also clarified which two peer feedback experiences has especially large heterogeneity in effect size: provided length and received length. The variation in effect size for provided length appeared to be predominantly quantitative, exhibiting generally positive effects but ranging from weak to strong across assignments. At the course level, the timing of the assignment (i.e., assignment #) within a course does not significantly influence the effect size of provided length. When examining variation at the assignment level, it's worth noting that the number of dimensions emerged as a significant moderating factor. Specifically, the effectiveness of providing length begins to diminish once the number of dimensions exceeds seven. This could be attributed to the increase in workload (i.e., more dimensions), which may affect student motivation (Akhteh et al., 2022) and, in turn, lead to a diminished effect. Additionally, the number of reviews is another potential factor that affects students' motivation. The combination of the number of reviews and dimensions within assignment could explain the observed effects. Conversely, received length involved qualitative variation, with effects sometimes being positive and at other times negative across different assignments. The social cognitive perspective (Bandura, 1989) provides a possible explanation for why the relationship with received length was sometimes negative: self-efficacy may be lowered by receiving large amount of critical feedback, which then influences students' efforts and outcomes (Schunk & DiBenedetto, 2020). However, because motivational levels were not directly observed, future research is required to directly test this explanation.

5.3. RQ3. To what extent is heterogeneity in effect size associated with content overlap between consecutive assignments

Importantly, the current study offers a novel perspective on understanding the considerable heterogeneity across assignments, demonstrating that the variations effect sizes of two different peer feedback

experience measures (i.e., provided length and received length) can be empirically explained by knowledge transfer. In general, the learning benefits of these two experiences tend to decrease as knowledge transfer becomes increasingly further, with the provided length effect size experiencing a smaller decline than the received length effect size (see Figs. 6 and 7). This pattern is consistent with social cognitive theory and the ICAP framework. Specifically, influences in the environment—such as assignments become novel and thus more challenging—facilitate targeted change in students' self-efficacy, outcome expectancies, or both. This process may lead to reductions in their beliefs about their abilities and their expectations for success, which could explain why gains in self-efficacy from these two peer feedback experiences may be reduced as greater transfer levels are tested. In terms of ICAP framework, providing feedback (constructive learning) may yield more robust learning outcomes than receiving feedback (i.e., passive learning), resulting in smaller declines in learning gains for provided length compared to received length as knowledge transfer becomes increasingly further.

It is also important to note that provided length demonstrated higher learning gains in near transfer, while received length was associated with very near transfer (see Figs. 6 and 7), which is consistent with ICAP framework and previous research. Similarly, constructive learning may result in more robust learning, and thus its learning benefits are likely to be observed at a relatively greater knowledge transfer levels than passive learning. As a related explanation, previous research has found that providing feedback has both direct and mediated pathways to learning through supporting revisions (Wu & Schunn, 2021, 2023) whereas, receiving feedback rarely contributes to learning unless it is accompanied by subsequent revisions (Wu & Schunn, 2023).

The relationship between transfer level and assignment genre change deserves further discussion. As depicted in Fig. 3, these two conceptualizations of content overlap were surprisingly independent: the distribution of transfer levels varies considerably between consecutive identical assignments and as well across substantially novel assignment genres (i.e., assignments vary in genre or topic). However, "near transfer" was found to occur in approximately 70% of cases when students worked on different drafts of the same assignment. This result was expected since the knowledge, skills and attitudes acquired from first draft can be readily applied to the next draft. Moreover, the distribution between near and far transfer changed to a 4:6 ratio in new writing assignments of various types. Just as salient though, very far transfer cases still occurred in every situation across genre-based categories. Motivational explanations could be relevant here: If students perform exceptionally well in one assignment, their motivation might decrease in the next assignment, especially when transitioning from the first draft to the subsequent draft (i.e., there is little need to improve). Finally, it was also surprising that the frequency of very near transfer cases were almost identical in the cases of different assignments. This could be attributed to the instructor creating sequences of assignments that built upon knowledge obtained in the prior assignments even when the genre changed. It might also reflect some stability in relative engagement of students: some students were more inclined to spend additional time trying to complete each assignment as well as contribute more to reviewing.

5.4. Practical implications

Several practical implications can tentatively be derived from the findings. First, because providing feedback appears to have a larger impact on learning than merely receiving feedback, instructors are encouraged to create conditions (e.g., by provide guidance and training to students) that encourage all students, not just the strong students, to provide detailed feedback, such as including clarifying expectations, offering examples, and emphasizing specific aspects of their work that can be improved (Nicol & Macfarlane-Dick, 2006; Sluijsmans et al., 2002). Furthermore, some web-based peer feedback systems allow instructors to set a minimum length for submitted comments or minimum number of provided comments.

Second, it is essential to thoughtfully consider the nature and progression of assignments when incorporating peer feedback activities into the curriculum. As noted earlier, knowledge transfer is closely associated with the effectiveness of peer feedback experiences. Therefore, from a design perspective, assignments should be logically structured and sequentially build upon one other to facilitate the transfer of knowledge, skills, and attitudes from one assignment to the next.

5.5. Limitations and future research

There are several limitations that should be noted in the present study. First, it is important to note that the statistical technique (i.e., regression analysis) used in the current study to examine the role of each peer feedback experience in shaping learning is fundamentally a correlation method. As a result, strong causal claims cannot be made. However, the regression models used time-series data which preclude reverse causality as well as included important controls related to most likely confounds based upon prior research, to increase the likelihood that effects on learning were estimated accurately. Further, the high-level finding of greater benefits of providing than receiving is consistent with prior experimental research (K. Cho & MacArthur, 2011; Ion et al., 2019; Lundstrom & Baker, 2009).

Second, quality of the peer feedback in our study is measured by length (i.e., number of words) rather than by direct evaluation of feedback features. We selected this approach for its ease of implementation across large-scale and cross-context datasets. Nevertheless, it is important to note that feedback features can exhibit substantial variation, even within comments of approximately the same length. For instance, some might contain elaboration and justification (Noroozi et al., 2016), or offer solution and mitigating praise (Wu & Schunn, 2020a). These different feedback features also shape the learning opportunities for both feedback providers and receivers. Given the considerable heterogeneity observed in this study concerning the effect sizes of provided length, future research could employ complex Natural Language Processing (NLP) techniques to automatically detect these features (Bauer et al., 2023; Darvishi et al., 2022), and further explore which specific features contribute most to learning within peer feedback.

Third, data collected from a single peer feedback system may have limited generalizability to other systems and to settings not using a web-based peer feedback tool. Using data from one system ruled out the confounding influence of the tool itself and the system is relatively similar to many other current web-based tools, but the potential impact of minor differences between various peer feedback systems should be considered. For instance, some peer feedback systems may impose a specific structure or format for feedback, such as rating scales, rubrics, or templates, while others allow more open-ended responses. Additionally, the implementation of backward evaluation also varies somewhat between different platforms, such as scales, comments, likes, and flags (Misiejuk & Wasson, 2021). Future research could address these limitations by extending beyond a single peer feedback system to formally test the generalizability of findings.

Further, although our exploratory analysis identified some factors associated with effect size heterogeneity, a substantial portion of unexplained heterogeneity remained. Future research could explore more micro-level details (e.g., similarity of comment prompts and rating dimensions across consecutive assignments; whether comment prompts focused reviewers on few versus many performance dimensions) or contextual factors (e.g., relative timing between assignments) to gain a more comprehensive understanding of the sources of heterogeneity.

6. Conclusion

On the whole, our study provides a comprehensive understanding of peer feedback benefits through multi-assignment datasets and learning

analysis methods. It differentiates the effects of providing and receiving feedback, reveals important variation in their effect sizes, and provide new insights into the role of knowledge transfer. These findings lead to several practical recommendations for instructors: 1) Providing feedback contributed more to performance and learning, and thus instructors are encouraged to employ multiple strategies across various aspects to encourage students to provide longer comments. This could be achieved through instructional support (e.g., training and guidance), system settings (e.g., minimum amount and length), and assignment components (e.g., comment prompt); 2) Given that the unique contributions of quantity in both providing and receiving feedback appear to be influenced by knowledge transfer, it would be beneficial for instructors to strategically design and implement consecutive assignments within course depending on students' existing knowledge, skills, and attitudes.

Credit author statement

Qiuchen Yu: Data curation, Formal analysis, Software, Visualization, Writing-original draft; Christian D. Schunn: Conceptualization, Formal analysis, Methodology, Resources, Supervision, Writing-review & editing.

Funding

This work was partially supported by the China Scholarship Council [Grant Number 202206770017].

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The co-author is a co-inventor of the peer review system used in the study-Christian D. Schunn.

Data availability

We have shared the research data and code in the supplementary material.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2023.107857.

Appendix A

Table A1
Within each context, number of assignments with mean and standard error for effect sizes for $Z\text{-}Score_{i,j}$, as well as the heterogeneity of corresponding effects within each context. Statistically significant moderation by context is indicated in bold.

Contextual factors	*n*	Mean	SE	I^2	*p*
Institution type					
High school	46	0.29	0.003	11%	.336
University	459	0.27	0.025	65%	<.001
Discipline					
STEM	317	0.28	0.020	61%	<.001
English	83	0.25	0.023	46%	<.001
Other	105	0.25	0.040	70%	<.001
Course size					
25–50[a]	189	0.31	0.037	52%	<.001
50–100[b]	175	0.27	0.024	56%	<.001
100– [c]	141	0.24	0.015	68%	<.001
Assignment#					
1st	230	0.29	0.031	67%	<.001
2nd–3rd	175	0.26	0.016	53%	<.001
4th–	100	0.24	0.020	57%	<.001
# of dimensions					
1–3[b]	95	0.22	0.022	55%	<.001
4–6[a]	260	0.28	0.024	63%	<.001
7–[a]	150	0.28	0.024	63%	<.001
Reviewing type					
One-at-a-time[b]	268	0.23	0.021	52%	<.001
All-at-once[a]	237	0.31	0.031	66%	<.001

Note. Different letters on the same subgroup indicate significant differences among groups at $\alpha = 0.05$.

Table A2
Within each context, number of assignments, with mean and standard error of effect sizes for $Provided\ Length_{i,j}$, as well as the heterogeneity of corresponding effects within each context. Statistically significant moderation by context is indicated in bold.

Contextual factors	*n*	Mean	SE	I^2	*p*
Institution type					
High school	46	0.11	0.050	100%	<.001
University	459	0.11	0.022	100%	<.001
Discipline					
STEM	317	0.12	0.020	100%	<.001
English	83	0.09	0.040	100%	<.001
Other	105	0.11	0.027	100%	<.001
Course size					

(*continued on next page*)

Table A2 (*continued*)

Contextual factors	*n*	Mean	SE	I^2	*p*
25–50	189	0.11	0.034	100%	<.001
50–100	175	0.11	0.027	100%	<.001
100–	141	0.13	0.010	100%	<.001
Assignment#					
1st	230	0.10	0.024	100%	<.001
2nd–3rd	175	0.13	0.020	100%	<.001
4th–	100	0.10	0.034	100%	<.001
# of dimensions					
1–3	95	0.10	0.035	100%	<.001
4–6[a]	260	0.13	0.023	100%	<.001
7–[b]	150	0.09	0.021	100%	<.001
Reviewing type					
One-at-a-time	268	0.13	0.019	100%	<.001
All-at-once	237	0.10	0.020	100%	<.001

Note. Different letters on the same subgroup indicate significant differences among groups at $\alpha = 0.05$.

Table A3
Within each context, number of assignments, with mean and standard error of effect sizes for *Received Length*$_{i,J}$, as well as the heterogeneity of corresponding effects within each context.

Contextual factors	*n*	Mean	SE	I^2	*p*
Institution type					
High school	46	−0.05	0.051	100%	<.001
University	459	−0.001	0.030	100%	<.001
Discipline					
STEM	317	−0.004	0.029	100%	<.001
English	83	−0.02	0.041	100%	<.001
Other	105	0	0.035	100%	<.001
Course size					
25–50	189	−0.008	0.052	100%	<.001
50–100	175	−0.002	0.030	100%	<.001
100–	141	−0.01	0.010	100%	<.001
Assignment#					
1st	230	−0.018	0.031	100%	<.001
2nd–3rd	175	0.001	0.028	100%	<.001
4th–	100	0.01	0.042	100%	<.001
# of dimensions					
1–3	95	−0.02	0.042	100%	<.001
4–6	260	−0.01	0.028	100%	<.001
7-	150	0.003	0.033	100%	<.001
Reviewing type					
One-at-a-time	268	−0.01	0.015	100%	<.001
All-at-once	237	0	0.024	100%	<.001

Table A4
Linear correlation coefficients between peer feedback experiences' effect sizes and the assignment-level statistical properties of that predictor.

	ES Z-Score$_{i,J}$	*ES Provided Length*$_{i,J}$	*ES Provided Helpfulness*$_{i,J}$	*ES Received Length*$_{i,J}$	*ES Received Helpfulness*$_{i,J}$
M *Z-Score*$_{i,J}$	[a]				
SD *Z-Score*$_{i,J}$	[a]				
Skew *Z-Score*$_{i,J}$	0.049				
Kurt *Z-Score*$_{i,J}$	−0.110*				
M *Provided Length*$_{i,J}$		−0.047			
SD *Provided Length*$_{i,J}$		−0.042			
Skew *Provided Length*$_{i,J}$		−0.085			
Kurt *Provided Length*$_{i,J}$		−0.099*			
M *Provided Helpfulness*$_{i,J}$			−0.046		
SD *Provided Helpfulness*$_{i,J}$			0.015		
Skew *Provided Helpfulness*$_{i,J}$			0.010		
Kurt *Provided Helpfulness*$_{i,J}$			0.044		
M *Received Length*$_{i,J}$				−0.078	
SD *Received Length*$_{i,J}$				−0.058	
Skew *Received Length*$_{i,J}$				−0.019	
Kurt *Received Length*$_{i,J}$				−0.081	
M *Received Helpfulness*$_{i,J}$					0.026
SD *Received Helpfulness*$_{i,J}$					−0.008
Skew *Received Helpfulness*$_{i,J}$					−0.001
Kurt *Received Helpfulness*$_{i,J}$					−0.003

Notes. M = mean; SD = standard deviation; Skew = skewness; Kurt = kurtosis. [a] correlation was not found since the Z-scores were normalized in each assignment. *$p <$ .05.

References

Akhteh, M. P., Farrokhnia, M., Banihashem, S. K., & Noroozi, O. (2022). The relationship between students' satisfaction and motivation and their perceived learning outcome in an online peer feedback module. In O. Noroozi, & I. Sahin (Eds.), *Studies on education, science, and technology 2022* (pp. 297–310). ISTES Organization.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*(9), 1175–1184. https://doi.org/10.1037/0003-066x.44.9.1175

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

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

Berggren, J. (2015). Learning from giving feedback: A study of secondary-level students. *ELT Journal, 69*(1), 58–70. https://doi.org/10.1093/elt/ccu036

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290–298. https://doi.org/10.1037/a0031026

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

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

Cho, Y. H., & Cho, K. (2010). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629–643. https://doi.org/10.1007/s11251-010-9146-1

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338. https://doi.org/10.1016/j.learninstruc.2009.08.006

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

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education, 48*(3), 409–426. https://doi.org/10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology, 53*(4), 844–875. https://doi.org/10.1111/bjet.13233

Davey, K. R. (2011). Student peer assessment: Research findings from a case study in a master of chemical engineering coursework-program. *Education for Chemical Engineers, 6*(4), e122–e131. https://doi.org/10.1016/j.ece.2011.08.004

Dong, Z., Gao, Y., & Schunn, C. D. (2023). Assessing students' peer feedback literacy in writing: Scale development and validation. *Assessment & Evaluation in Higher Education, 1–16*. https://doi.org/10.1080/02602938.2023.2175781

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Dunlap, J. C., & Grabinger, S. (2003). Preparing students for lifelong learning: A review of instructional features and teaching methodologies. *Performance Improvement Quarterly, 16*(2), 6–25. https://doi.org/10.1111/j.1937-8327.2003.tb00276.x

Ertmer, P. A., Richardson, J. C., Belland, B., Camin, D., Connolly, P., Coulthard, G., Lei, K., & Mong, C. (2007). Using peer feedback to enhance the quality of student online postings: An exploratory study. *Journal of Computer-Mediated Communication, 12*(2), 412–433. https://doi.org/10.1111/j.1083-6101.2007.00331.x

Gamage, D., Staubitz, T., & Whiting, M. (2021). Peer assessment in MOOCs: Systematic literature review. *Distance Education, 42*(2), 268–289. https://doi.org/10.1080/01587919.2021.1911626

Gao, Y., An, Q., & Schunn, C. D. (2023). The bilateral benefits of providing and receiving peer feedback in academic writing across varying L2 proficiency. *Studies In Educational Evaluation, 77*, Article 101252. https://doi.org/10.1016/j.stueduc.2023.101252

Gao, Y., Schunn, C. D. D., & Yu, Q. (2019). The alignment of written peer feedback with draft problems and its impact on revision in peer assessment. *Assessment & Evaluation in Higher Education, 44*(2), 294–308. https://doi.org/10.1080/02602938.2018.1499075

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Hardavella, G., Aamli-Gaagnat, A., Saad, N., Rousalova, I., & Sreter, K. B. (2017). How to give and receive feedback effectively. *Breathe, 13*(4), 327–333. https://doi.org/10.1183/20734735.009917

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2018). Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment & Evaluation in Higher Education, 43*(6), 955–968. https://doi.org/10.1080/02602938.2018.1424318

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

Ion, G., Sánchez Martí, A., & Agud Morell, I. (2019). Giving or receiving feedback: Which is more beneficial to students' learning? *Assessment & Evaluation in Higher Education, 44*(1), 124–138. https://doi.org/10.1080/02602938.2018.1484881

Kasch, J., Van Rosmalen, P., Henderikx, M., & Kalz, M. (2022). The factor structure of the peer-feedback orientation scale (PFOS): Toward a measure for assessing students' peer-feedback dispositions. *Assessment & Evaluation in Higher Education, 47*(1), 15–28. https://doi.org/10.1080/02602938.2021.1893650

Kerman, N. T., Noroozi, O., Banihashem, S. K., Karami, M., & Biemans, H. J. A. (2022). Online peer feedback patterns of success and failure in argumentative essay writing. *Interactive Learning Environments*, 1–13. https://doi.org/10.1080/10494820.2022.2093914

Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods, 10*(1), 83–98. https://doi.org/10.1002/jrsm.1316

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125–141. https://doi.org/10.1207/s15326985ep3502_6

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. https://doi.org/10.1080/08957347.2021.1933980

Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290. https://doi.org/10.1080/13562510600680582

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

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

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

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education, 19*(4), 209–232. https://doi.org/10.1080/08993400903384844

Lv, X., Ren, W., & Xie, Y. (2021). The effects of online feedback on ESL/EFL writing: A meta-analysis. *The Asia-Pacific Education Researcher, 30*(6), 643–653. https://doi.org/10.1007/s40299-021-00594-6

Martin, I. A., & Sippel, L. (2021). Providing vs. receiving peer feedback: Learners' beliefs and experiences. *Language Teaching Research*. https://doi.org/10.1177/13621688211024365, 136216882110243.

Misiejuk, K., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education, 175*, 104319. https://doi.org/10.1016/j.compedu.2021.104319.

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior, 117*, Article 106658. https://doi.org/10.1016/j.chb.2020.106658

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*(4), 375–401. https://doi.org/10.1007/s11251-008-9053-x

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Noroozi, O. (2022). The role of students' epistemic beliefs for their argumentation performance in higher education. *Innovations in Education & Teaching International, 1–12*. https://doi.org/10.1080/14703297.2022.2092188

Noroozi, O., Banihashem, S. K., Taghizadeh Kerman, N., Parvaneh Akhteh Khaneh, M., Babayi, M., Ashrafi, H., & Biemans, H. J. A. (2022). Gender differences in students' argumentative essay writing, peer review performance and uptake in online learning environments. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2022.2034887

Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education, 31*, 20–31. https://doi.org/10.1016/j.iheduc.2016.05.002

Noroozi, O., Hatami, J., Bayat, A., van Ginkel, S., Biemans, H. J. A., & Mulder, M. (2018). Students' online argumentative peer feedback, essay writing, and content learning: Does gender matter? *Interactive Learning Environments, 28*(6), 698–712. https://doi.org/10.1080/10494820.2018.1543200

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. https://doi.org/10.1007/s11251-015-9353-x

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. https://doi.org/10.1080/03075079.2017.1320374

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology, 108*(8), 1098–1120. https://doi.org/10.1037/edu0000103

Philippakos, Z. A., & MacArthur, C. A. (2016). The effects of giving feedback on the persuasive writing of fourth- and fifth-grade students. *Reading Research Quarterly, 51*(4), 419–433. https://doi.org/10.1002/rrq.149

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

Potter, T., Englund, L., Charbonneau, J., MacLean, M. T., Newell, J., & Roll, I. (2017). ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry, 5*(2), 89. https://doi.org/10.20343/teachlearninqu.5.2.8

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066. https://doi.org/10.1037/edu0000190

Schunk, D. H., & DiBenedetto, M. K. (2020). Motivation and social cognitive theory. *Contemporary Educational Psychology, 60*, Article 101832. https://doi.org/10.1016/j.cedpsych.2019.101832

Schunn, C. (2016). Writing to learn and learning to write through sword. In *Adaptive educational technologies for literacy instruction* (pp. 243–260). Routledge. https://doi.org/10.4324/9781315647500-16.

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. https://doi.org/10.1002/jaal.525

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

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *International Review of Research in Open and Distance Learning, 15*(3). https://doi.org/10.19173/irrodl.v15i3.1680

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Tsivitanidou, O. E., Zacharia, Z. C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction, 21*(4), 506–519. https://doi.org/10.1016/j.learninstruc.2010.08.002

Vickerman, P. (2009). Student perspectives on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*(2), 221–230. https://doi.org/10.1080/02602930801955986

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

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

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

Wu, Y., & Schunn, C. D. (2020b). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology, 62*, Article 101897. https://doi.org/10.1016/j.cedpsych.2020.101897

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

Wu, Y., & Schunn, C. D. (2023). Passive, active, and constructive engagement with peer feedback: A revised model of learning from peer feedback. *Contemporary Educational Psychology, 73*, Article 102160. https://doi.org/10.1016/j.cedpsych.2023.102160

Xu, Q., & Peng, H. (2022). Exploring learner motivation and mobile-assisted peer feedback in a business English speaking course. *Journal of Computer Assisted Learning, 38*(4), 1033–1045. https://doi.org/10.1111/jcal.12660

Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education, 47*(7), 1116–1128. https://doi.org/10.1080/02602938.2021.2001431

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review, 37*, Article 100484. https://doi.org/10.1016/j.edurev.2022.100484

Yuan, J., & Kim, C. (2015). Effective feedback design using free technologies. *Journal of Educational Computing Research, 52*(3), 408–434. https://doi.org/10.1177/0735633115571929

Zhang, F., Schunn, C., Li, W., & Long, M. (2020). Changes in the reliability and validity of peer assessment across the college years. *Assessment & Evaluation in Higher Education, 45*(8), 1073–1087. https://doi.org/10.1080/02602938.2020.1724260

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

Zong, Z., Schunn, C. D., & Wang, Y. (2021b). What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior, 124*, Article 106924. https://doi.org/10.1016/j.chb.2021.106924