Instructional Science
https://doi.org/10.1007/s11251-023-09640-w

ORIGINAL RESEARCH

When do students provide more peer feedback? The roles of performance and prior feedback experiences

Zheng Zong[1,2] · Christian D. Schunn[3] · Yanqing Wang[2]

Received: 13 February 2021 / Accepted: 25 June 2023

Abstract
Students benefit from receiving and providing peer feedback, but the degree of participation limits the benefit. Further, students sometimes resist participation, providing few or only short comments. Prior researchers have examined the role of general attitudes toward peer feedback in limiting participation. However, little research has examined how peer feedback experiences predict the subsequent amount of feedback that students provide to peers. Data on peer feedback experiences and behaviors across multiple assignments were taken from students across two psychology courses (*N=360*), two biology courses (*N=483*), and one astronomy course (*N=170*). The zero-inflated negative binomial (ZINB) regression analyses reveal that receiving fewer critical peer comments in the prior assignment, recognition for higher quality feedback in the prior assignment, and stronger performance on the current assignment predicted higher participation in peer feedback, but norm-setting did not appear to have a role. Implications for practitioners are discussed.

Keywords Peer assessment · Peer feedback · Peer recognition · Participation in peer feedback

Introduction

Peer review is a teaching method that can address insufficient teaching resources in higher education while also using an effective pedagogical technique (Topping, 1998). As a result, peer review has become an indispensable part of instruction using complex artifacts, such as essays (Applebee & Langer, 2011; Valero Haro et al., 2019), video presentations (Min,

✉ Yanqing Wang
yanqing@hit.edu.cn

[1] School of Law and Economics, China University of Political Science and Law, Beijing 100088, China

[2] School of Management, Harbin Institute of Technology, 13 Fayuan St., Nangang District, Harbin 150001, China

[3] Learning Research & Development Center, University of Pittsburgh, PAPennsylvania 15260, USA

Published online: 14 July 2023

2016), design projects (Rodriguez et al., 2018; Whicher et al., 2018), or computer code (Ballantyne et al., 2002; Wang et al., 2012). Peer review is often called *peer assessment* when focusing on the peers' numerical ratings and *peer feedback* when focusing on the peers' comments. Participating in peer feedback is generally found to be beneficial for students (van Zundert et al., 2010; Chang, 2016; Rotsaert et al., 2018; Zong et al., 2021a).

The validity or reliability of peer assessment has been an essential line of peer review research for over two decades (Haaga, 1993; Stefani, 1994; Marcoulides & Simkin, 1995; Mowl & Pain, 1995; Cheng & Warren, 1999; Cho et al., 2006; Cho & MacArthur, 2011; Schunn et al., 2016). A meta-analysis of these studies revealed an adequate average validity of peer assessments (Li et al., 2016). High reliability and validity were consistently found when peer assessment is: supported by a rubric, done online rather than on paper, and combined as ratings and comments rather than ratings alone. There has also been significant research examining the learning benefits of participating in peer feedback (e.g., Cho and Schunn, 2007; Lundstrom and Baker, 2009; van Zundert et al., 2010; Crinon, 2012; Adachi et al., 2018; Deiglmayr, 2018; Huisman et al., 2018; Martin and Evans, 2018; Zong et al., 2021a). A recent meta-analysis of the learning benefits of peer feedback (Li et al., 2020) revealed positive effects on average and strong learning benefits when: using a rubric, completed online, done anonymously, and implemented in higher education. Peer assessment with these characteristics is becoming increasingly popular through support from several widely available tools, such as *CPR* (calibrated peer review, http://cpr.molsci.ucla.edu), *Peerceptiv* (https://peerceptiv.com), *Kritik* (https://kritik.io), *EliReview* (https://elireview.com), *EduFlow* (https://eduflow.com), and *FeedbackFruits* (https://feedbackfruits.com).

However, all of these positive outcomes depend upon students' willingness to participate actively in peer feedback and provide in-depth comments to their peers. Unfortunately, students sometimes provide no feedback or only minimal amounts of feedback (Liu & Carless, 2006; Feng et al., 2019). The greater the amount of peer feedback provided by students, the more the feedback recipients revise (Popp & Goldman, 2016; Zhang et al., 2017; Wu & Schunn, 2021), and the more the providers improve their writing (Cho & MacArthur, 2011; Wu & Schunn, 2021).

Students' attitudes toward peer assessment are often mixed (Zou et al., 2018), which can limit their participation in the peer feedback process (van Zundert et al., 2010; Chang, 2016). Because of the relatively low level of their peers' current expertise, students usually do not think that their grades should be determined by their peers (Kaufman & Schunn, 2011), or they worry that the feedback they receive will not meet their curriculum needs (Mangelsdorf, 1992; Liu & Carless, 2006). Previous studies have investigated students' concerns about motivation to participate and encouraged students to actively participate in peer evaluation by providing training (e.g., Sluijsmans et al., 2002) or providing information about peer assessment reliability (Jones & Alcock, 2014).

However, the existing literature rarely pays attention to students' actual experience with the peer feedback process. Some scholars suggest that such experience may ease negative attitudes caused by the simple lack of familiarity with peer feedback (Kankanhalli et al., 2005; Liu & Carless, 2006). Training does improve the validity of student feedback (Sluijsmans et al., 2002; Schunn et al., 2016). However, the claim that simple experiences with peer feedback will change participation is relatively untested, and it is unclear what kinds of experiences will be more critical (e.g., observing good models or being rewarded for fair reviewing). This study considers how experience with peer feedback shapes the willingness to provide feedback in later assignments. Once this relationship is established, it will be possible to create new interventions that will help to ensure stronger and more consistent participation in peer feedback.

Theoretical background

Experiences that lead to self-regulation in peer feedback. Repeated experience with a peer review system can be conceptualized as a feedback-practice cycle (Berg, 1999; Min, 2016) or a kind of agentic engagement with feedback (Winstone et al., 2017). Over peer review cycles, students can regularly observe peer comments, receive evaluations of their performance on the writing tasks, and receive evaluations of their reviewing performance, and each of these experiences can shape their future reviewing behaviors. Such experiences that iteratively guide changes in behavior can be conceptualized in terms of self-regulation, a framework that has previously been invoked as relevant to peer review (Harris et al., 2015; Brown et al., 2016; Meusen-Beekman et al., 2016). Self-regulation can involve three levels of change (Zimmerman & Risemberg, 1997; Dinsmore et al., 2008; Chen & Chiu, 2016): *personal regulation*, which refers to attitudinal regulation like changing self-efficacy; *behavioral regulation*, which relates to modifying task behaviors like changing levels of participation in providing peer feedback (e.g., providing more/fewer comments or longer/shorter comments); and *environmental regulation*, which involves seeking social supports and creating a more productive task environment. The three levels are thought to be interdependent. For example, changes in task behaviors can come about from self-efficacy changes, such as being less willing to participate in peer feedback with low self-efficacy for the feedback topic. Two of these levels are particularly relevant to student participation in peer feedback.

Within assignments involving peer review, students receive multiple sources of information that can shape their self-efficacy. Individuals with high self-efficacy for a domain are more likely than individuals with lower self-efficacy to complete tasks related to that domain. In surveys and interviews, students often mention concerns about their abilities in the domain as a reason for wanting to avoid participating in peer review (Liu & Carless, 2006; Kaufman & Schunn, 2011; Moore & Teather, 2013). Through peer review processes, students can receive task evaluations that further shape their self-efficacy (To & Panadero, 2019). For example, if a student receives many negative comments on one assignment, their self-efficacy for the general topic (or even the course overall) may decline. Moore and Benbasat (1991) found that when participants perceived their contributions to be relatively strong, they were more inclined to participate in community knowledge exchange. Thus, we hypothesize that students who receive negative evaluations (e.g., more negative comments or lower ratings) from each peer in a prior assignment might provide less peer feedback (fewer comments and shorter comments) to each peer in the next assignment because of a decrease in self-efficacy. An examination of the contents of a random selection of long comments from a stratified random sample of 40 students (sampled from extreme-group cases of very high and very low performing author and reviewer pairs) revealed that long comments (i.e., those having at least 40 words) almost universally included criticism (i.e., between 95% and 99% of comments, depending upon author-reviewer pairing). Therefore, the total number of comments is expected to be highly correlated with the number of negative comments. Similarly, when comments are negative, they tend to be longer, then the average length of comments might be substantially correlated with the negativity of comments. In a particular course, comments may tend towards negative if the reviews are anonymous (Rotsaert et al., 2018; Panadero & Alqassab, 2019) or if training directs students to attend to problems (Sluijsmans et al., 2002; Min, 2016; van Blankenstein et al., 2019). Thus, we proposed that:

H1: High task quality in the prior assignment will predict more active participation in peer review.

H2: More feedback (which is predominantly negative) in the prior assignment predicts less active participation in peer review.

Another factor that influences self-efficacy involves mastery experiences. In completing an assignment (that will later undergo peer review), students will have a sense of struggle or success (i.e., engage in an implicit self-assessment, Winstone et al., 2017), which then influences their self-efficacy for the underlying domain (Margolis & McCabe, 2003; Zeldin et al., 2008). That self-efficacy change would then similarly influence participation in the peer-reviewing phase of that assignment. Thus, we hypothesize that relative performance on the assignment undergoing peer review will predict the level of participation in peer feedback (number of comments and length of comments provided) in that assignment:

H3: Higher task quality in the current assignment will predict more active participation in peer review.

In addition to peer feedback connections to self-efficacy changes, there are connections between peer feedback to norm-setting effects (Flower et al., 1986; Abubakar et al., 2019) and reinforcement feedback (Daniels, 2016) on behavioral regulation. In general, organizations often rely on peer feedback to improve their members' performance through norm-setting and reinforcement feedback (VanStelle et al., 2012; Wang et al., 2017; Fernandes et al., 2019). From this more general research on peer feedback, we hypothesize that students can learn how much feedback is expected to be provided through the amount of feedback they receive from others as a kind of norm-setting (i.e., when receiving more or longer comments, they will provide more or longer comments in the next assignment). However, note that norms for a particular kind of activity tend to be set early on within a context, in this case, the ways in which peer review should be conducted (Kaufman & Schunn, 2011). Thus, norms may play little role in changes in the amount of feedback provided from one assignment to the next:

H4: Greater amounts of feedback received in the prior assignment (both amount and length) predicts more active participation in peer review.

We also hypothesize that peers' positive recognition of comments previously provided to peers serves as a kind of reinforcement feedback (i.e., when receiving praise for prior comments provided, they will provide more or longer comments in the next assignment). Rewarding students for more helpful feedback increased the length of comments that peers produced in their peer feedback (Patchan et al., 2018), but that study did not examine change over time with experience. In research on peer feedback on computer programs, Zong et al. (2022) found that when students were recognized for especially good reviews in a prior assignment, they tended to provide longer feedback to peers on the next assignment, but only if they were poorer-performing students. However, the particular system had quite infrequent feedback on this dimension from students. A more common approach across online peer feedback systems is to have all received comments rated for helpfulness.

A more robust effect could be observed with this more in-depth and consistent feedback to reviewers. Thus, we hypothesized that:

H5: Higher helpfulness ratings from peers for provided feedback in the prior assignment will predict more active participation in peer review.

In sum, we hypothesize that students learn to regulate their degree of participation in peer feedback through three sources (see Fig. 1): (1) updates to self-efficacy via feedback or mastery experiences with task success in the prior or current assignment; (2) norm-setting via the amount of feedback they receive, and (3) reinforcement messages regarding the value of the prior feedback provided to their peers. To test these three hypotheses, we conduct analyses of gradual growth and decline in the amount of feedback provided from one assignment to the next due to their prior and current assignment experiences.

Dimensions of behavioral regulation in peer feedback. As noted earlier, we make hypotheses about two aspects of participation in peer feedback: the number of comments provided and the length of comments provided. Why those two dimensions? The degree of participation in peer feedback can be conceptualized and measured in a variety of ways. Provided feedback can be of varying accuracy and varying impact on the feedback recipient. However, the feedback provider can only indirectly control those aspects. For example, a student could try to participate to a greater extent but still not provide more impactful or accurate feedback because of a lack of knowledge. Alternatively, the degree of participation could be conceptualized in more simple quantitative terms, and this dimension of participation is more directly under the feedback provider's control. We, therefore, test these hypotheses at this more quantitative level. In particular, we conceptualize the degree of participation in terms of the number of comments provided to peers (Min, 2005, 2006; Zong et al., 2021b) as well as the length of comments provided (Hamer et al., 2015; Paltridge, 2015; Patchan et al., 2018). The norm-setting measures will consist of the number of comments received and the length of comments received for parallel form.

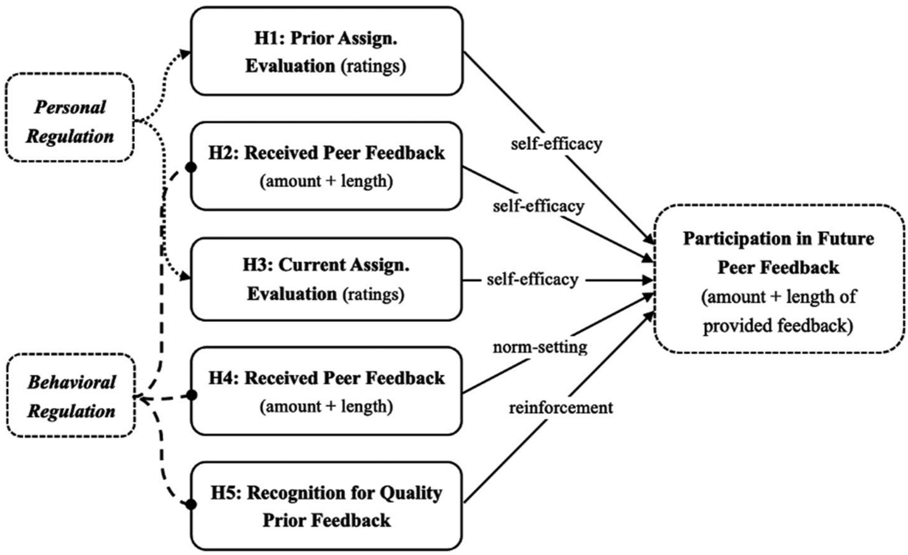

Fig. 1 Hypothesized effects of personal regulation and behavioral regulation on the amount of participation in peer feedback

Methods

Course setting and participants

Peer feedback from 1,033 undergraduate students was examined: 360 students (74% female; 82% Caucasian, 10% African American, and 8% Asian; median age 21) across two offerings of a mid-level introductory psychology course entitled *Introduction to Cognitive Psychology*, 170 students (45% female; 95% Caucasian; median age 21) from a mid-level introductory astronomy course, entitled *Contemporary Astronomy*, and 483 students (59% female; 69% Asian, 18% Hispanic/Latino, 12% Caucasian; median age 21) across two offerings of a writing-intensive course for biology majors in biology entitled *Scientific Writing*. The two offerings of a given course were in different years.

The courses were from different universities, but all three universities were large research-oriented, selective public universities in the US. These three in-person courses were selected to test the generality of patterns in how students change their willingness to give peer feedback. These courses utilized the same online peer assessment system, *Peerceptiv*, as described below, to produce directly comparable measures of peer feedback performance. Among the many courses using *Peerceptiv*, these three were selected to represent various disciplines while focusing on larger courses (to increase statistical power) that also had at least four writing assignments (to study change across writing assignments). Based upon the number of courses using *Peerceptiv* at these universities and the common use of peer feedback in 1st year writing courses in US universities, it is likely that most students had some familiarity with face-to-face peer feedback but little prior experience with *Peerceptiv* in particular. Variations in prior experiences with peer feedback may shape initial variation in the amount and length of peer feedback provided.

The number of writing tasks varied across the courses, but all involved in-depth writing about scientific content from a particular discipline. The psychology courses involved six assignments with peer review: one year, two drafts each of three different writing assignments, and in the other year, it was one draft each of six different writing assignments. In both cases, the writing assignments were similar: choose from a fixed set of writing prompts that apply concepts from the course to specific situations. The astronomy course had one larger semester-long project involving data analysis and argumentation, divided into four parts, one writing assignment per part. The biology course involves five assignments, four of which are components of one larger research project: writing about post-graduation plans; describing a novel research question and hypothesis; summarizing two research articles related to the research question; describing a novel experiment designed to test the hypothesis; a longer paper integrating introduction through research methods.

Materials

Students did all reviewing activities via *Peerceptiv*, which is commonly used with writing-based assignments. The system has the following features, which were all used in the selected courses: (1) *asynchronous reviewing* (i.e., via forms, not interactive), (2) *reviewing including both numeric evaluation and open-response feedback comments*, (3) *double-blind reviewing* (i.e., reviewers and authors are anonymous), (4) *multiple reviews per document* (i.e., 4 to 5 reviews per document, provided and therefore also received), (5) dimension-specific reviewing prompts; and 5) *mechanisms in the system to encourage more constructive suggestions* (i.e., authors judge whether the feedback was helpful, which produces a grade for the reviewer; Patchan et al., 2018). These standard features are found in many online peer assessment systems (e.g., Calibrated Peer Review, Chapman and Fiore, 2000; ELI Peer Review, McLeod et al., 2013). Most saliently, for the current study, the students could submit a variable number of comments within each review, which is a critical variable in the current study. Figure 2 presents the student interface pages in *Peerceptiv*. Particularly important to this study is the presence of multiple textboxes per reviewing dimension for entering comments; students only had to submit one comment per dimension, but they could provide more comments.

Measures

The current study uses a large dataset automatically collected within *Peerceptiv*: hundreds of thousands of comments from tens of thousands of reviews. The analyses are based upon measures derived from this dataset that is organized by assignment (e.g., the total amount of feedback provided on one assignment across reviews). Thus, the dataset size for the regressions is the number of students multiplied by the number of assignments in the course minus one (because the first reviewing assignment serves as the baseline). Since the two courses in psychology and biology are relatively homogeneous (similar assignment types, similar reviewing focus, similar student populations) and the courses showed identical patterns in pilot data analyses, the data across both course offerings were combined into one dataset for each context. In terms of the number of comments produced after removing cases in which students did not submit a task or participate in a round: across both psychology courses, N = 1,506 produced by six assessments with 360 students; in the astronomy course, N = 508 produced by four assessments with 170 students; and across both biology courses, N = 1,920 produced by five assessments with 483 students.

The specific measures were defined as follows (see Table 1 for a summary and Table 2 for the maximum, means, and SDs of each variable in each course context). The outcome variable was defined based on the reviewing behavior on the J^{th} assignment, where J is the assignment number. Predictor variables were defined in terms of the $(J\text{-}1)^{th}$ assignment's reviewing behaviors and the J^{th} assignment's submission performance.

There were two participation measures in peer feedback serving as the study's primary outcome measures: the number of comments and the average length of comments.

Number of comments provided. In the system, students enter comments into text boxes for each reviewing dimension. There are multiple available textboxes for each dimension (as shown in Fig. 2). Thus, a student had to provide at least one comment per dimension but could give more than one comment. Collectively, across the documents they were assigned to review, reviewers could produce many more comments than the minimum required. For the time-series analysis, $\#Provided_J$ refers to the number of comments provided on the J^{th} assignment. This variable is calculated for assignment 2 through N as a dependent variable and for assignment *1* through *N-1* as the baseline predictor variable. Note that sometimes students addressed more than one topic (i.e., gave multiple comments) within one comment box. Given the very large size of the dataset, it was not feasible to manually segment all of the comments into idea units. A given student's tendency to include multiple ideas in one comment is regressed in the regression approach. Further, the inclusions of both the number

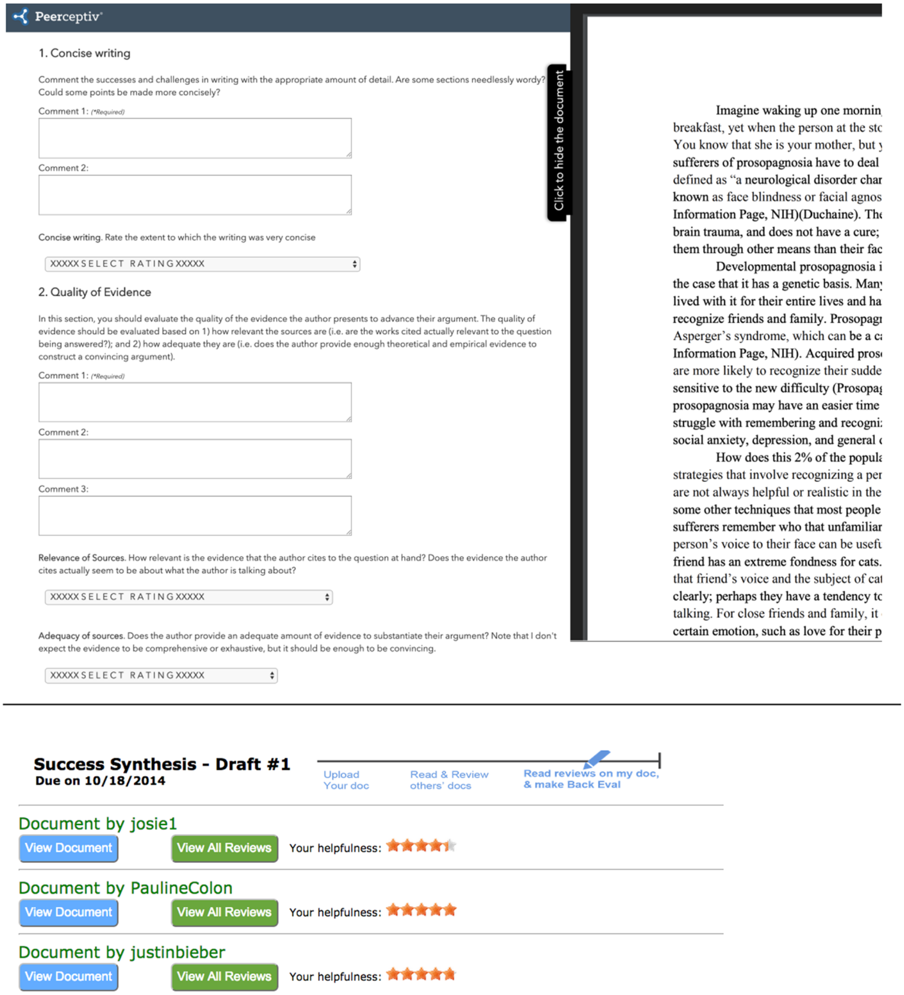

Fig. 2 Top: The major student reviewing interface within Peerceptiv. Students enter open-ended comments in the text boxes on the left organized by reviewing prompt and interwoven with drop-down ratings for each reviewing dimension. Bottom: The interface showing the summary of helpfulness ratings for each completed review (averaged across reviewing dimensions)

of comments and length of comments capture behavioral variation regardless of whether it occurs inside each textbox or across text boxes.

Length of comments provided. There was no standard minimum length for a given comment beyond needing to include at least one word. Reviewers could choose whether or not to identify a problem clearly, explain the problem's nature, provide suggested revisions, and praise the submission (overall or in some aspect). Given the complexity of the writing, one comment could involve a long paragraph. A mean comment length (in words) was calculated per comment provided across all dimensions and reviews on assignment J (*Length-Provided*$_J$). The length of comments provided is calculated for assignments 2

Table 1 Definition of each measurement variable, along with the connection to tested hypotheses

Measure	Definition	Hypothesis connection
Dependent Variables		
$\#Provided_J$	The number of peer feedback comments provided by a student across the completed reviews on the J^{th} assignment.	
$Length\text{-}Provided_J$	The mean number of words provided by a student per comment on the J^{th} assignment.	
Independent Variables		
$Low\text{-}Score_{J-1}$	*1* if the document score on the $(J\text{-}1)^{th}$ assignment is lower than the median score, and *0* otherwise (higher or not submitted)	H1: Self-efficacy
$\#Received_{J-1}$	The number of peer feedback comments received by a student on the $(J\text{-}1)^{th}$ assignment.	H2:Self-efficacy H4:Norm-setting
$Length\text{-}Received_{J-1}$	The mean number of words received by a student per comment on the $(J\text{-}1)^{th}$ assignment.	H2:Self-efficacy H4:Norm-setting
$\#Provided_{J-1}$	The number of peer feedback comments provided by a student across the completed reviews on the $(J\text{-}1)^{th}$ assignment.	
$Length\text{-}Provided_{J-1}$	The mean number of words provided by a student per comment on the $(J\text{-}1)^{th}$ assignment.	
$Z\text{-}Score_J$	The standardized score of the student's document on the J^{th} assignment (i.e., score minus mean assignment grade / SD of assignment grade, then divided into 5 levels according to the size of the value)	H3: Self-efficacy
$Recognition_{J-1}$	The mean helpfulness rating (Peerceptiv) received by a student for their reviewing on the $(J\text{-}1)^{th}$ assignment.	H5: Reinforcement
Course	Arbitrary indicator for a course section within a course discipline. *0* = 1st course, *1* = 2nd course	
J	The assignment number	

Table 2 Mean and standard deviations for each variable within each course, along with maximum observed values on each variable

Variable	Psychology			Astronomy			Biology		
	Max	Mean	SD	Max	Mean	SD	Max	Mean	SD
#Provided$_J$	55	21.7	11.1	48	19.2	9.9	120	22.8	10.6
Length-Provided$_J$	211	40.7	22.8	66	21.2	10.7	246	60.0	32.1
#Provided$_{J-1}$	55	22.1	10.6	48	17.6	8.8	120	20.1	12.0
Length-Provided$_{J-1}$	211	43.0	23.9	115	24.4	13.0	218	51.8	28.8
#Received$_{J-1}$	53	22.9	5.6	45	17.4	7.4	61	20.3	10.2
Length-Received$_{J-1}$	113	13.3	83.7	62	24.5	8.6	149	51.2	19.0
Recognition$_{J-1}$	5.0	4.4	0.5	5.0	3.6	1.0	5.0	4.0	0.6
Low-Score$_{J-1}$	1.0	0.5	0.5	1.0	0.4	0.5	1.0	0.4	0.5
Z-Score$_J$	4.0	2.2	1.4	4.0	2.2	1.4	4.0	2.1	1.4

through *N* as a dependent variable and for assignments 1 through *N-1* as the baseline predictor variable.

There was a range of predictor variables related to the hypotheses as well as several control variables important to addressing confounds and allowing for the combination of assignments into a single regression.

Number of comments received. In a complementary way, students also received a varying number of feedback comments on a given assignment submission, summed across reviewing dimensions and from multiple reviewers. If no document was submitted, the value was treated as missing. For the time-series analysis, $\#Received_J$ refers to the number of comments provided on the J^{th} assignment.

Length of comments received. As a possible source of change, a mean comment length (in words) was also calculated for reviews received ($Length\text{-}Received_J$) across all dimensions and reviews on assignment *J*. If no document was submitted for review on a given assignment, the value was treated as missing.

Recognition for helpful feedback. The *Peerceptiv* system has a mechanism by which each student, as a document author, evaluates the helpfulness of the comments they received on their document, called back evaluations, which can be thought of as direct performance feedback on the quality of the provided peer comments. In particular, each document's author rated helpfulness on a *1* to *5* scale. A mean rating was calculated across all provided reviews for each assignment *J* and was called $Recognition_J$. When authors failed to complete the back-evaluation step for a given review, the value was treated as missing, and a mean was calculated based on available data.

Low prior assignment score. Through the multi-peer feedback process, students receive an assessment of their mastery of course content and enhance their ability to write about course content, shaping their confidence in providing feedback to peers. Assessments based on detailed rubrics and multiple peers are likely to be valid and reliable (Li et al., 2016), particularly when given vital rubrics and incentives to seriously take the task (Cho & Schunn, 2007; Patchan et al., 2018). However, the students' memory of their exact scores for a specific prior assignment is likely to be imperfect, especially by the time they are reviewing for the next assignment (potentially three weeks or more later). Thus, for this study, relative performance on the prior assignment was treated as a categorical indicator variable instead of a continuous measure. In particular, $Low\text{-}Score_{J-1}$ was set to *1* if the score on the $(J-1)^{th}$ assignment was below the class median and *0* otherwise. Moreover, the use of a categorical indicator reduces multicollinearity problems with the measure of current assignment performance.

Current assignment score. Although students will not have received performance feedback on the current assignment by the time they are reviewing, they likely would have a sense of their overall performance, perhaps from how long it took to complete the assignment or how often they felt confused. Further, their mastery of the current assignment will likely shape how easily they can find problems and suggest solutions. Thus, the relative score ($Z\text{-}Score_J$) from the peer ratings on each given assignment was calculated as follows: (score obtained by the reviewer on assignment *J* – mean score for assignment *J*) / standard deviation of scores for assignment *J*. It was treated as missing if there was no submission for the current assignment.

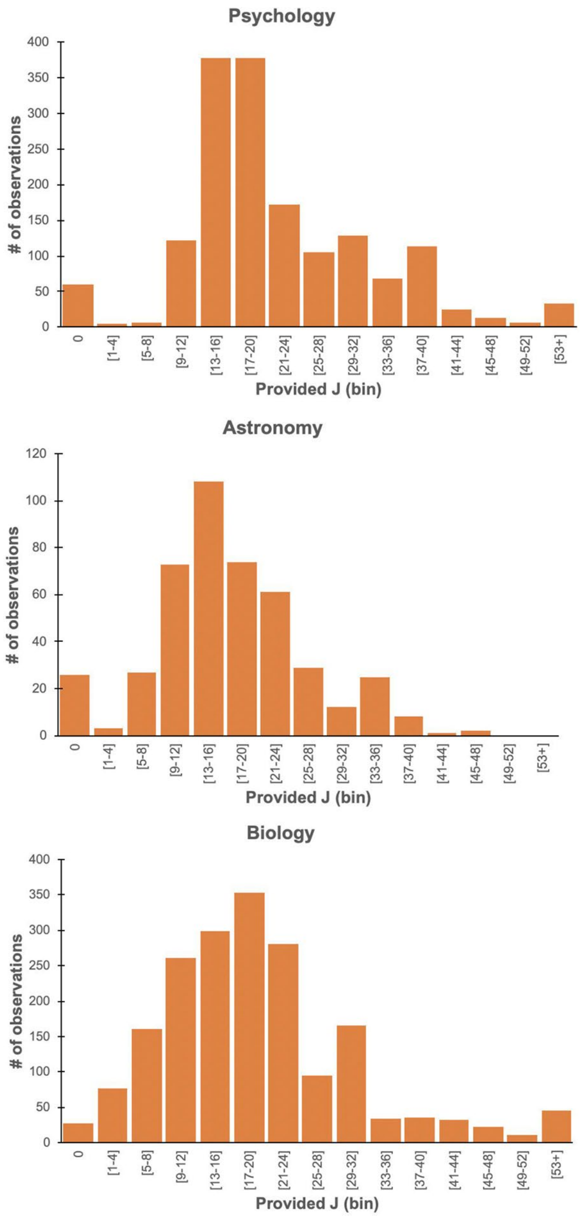

Fig. 3 Frequency histograms for the number of comments provided on the J^{th} assignment within each course

Assignment round. To account for general temporal trends in reviewing behaviors, variable *J* was the reviewing assignment number (2 to 6 in psychology, 2 to 4 in astronomy, and 2 to 5 in biology).

Course. It accounts for small differences in mean values in the variables across the two-course offerings within the psychology and biology courses. A *Course* indicator variable was created: set to *0* for the first-course offering and *1* for the second-course offering.

Analyses

The closest non-outlier value replaced outlier values in each continuous predictor variable. Such outliers occurred for at most 0.4% of values on any given variable and less than 0.1% of values on most variables.

The overall analytic approach was to use multiple regression, with $Provided_J$ and $Length\text{-}provided_J$ as the dependent variables respectively, $Provided_{J-1}$ and $Length\text{-}provided_{J-1}$ as the baseline control respectively (Barnett et al., 2005), $Received_{J-1}$, $Recognition_{J-1}$, $Low\text{-}Score_{J-1}$, and $Z\text{-}Score_J$ as core predictors, and *N* and *Course* as additional control variables. Separate regressions ran for each course context: (psychology courses vs. astronomy course vs. biology courses) given the substantial differences in student background and objects being reviewed.

Since the dependent variables, the number of feedback comments and the number of words per comment provided by a student, are count variables, they had non-normal distributions with a large right (positive) skew. Thus, traditional linear regression was not the best modeling method. For such positive skew cases involving count data, either Poisson regression or negative binomial regression distribution is recommended. Poisson is the appropriate choice when the outcome variable mean and variance (SD squared) are equal (Grogger & Carson, 1991; Gardner et al., 1995). Since the variance for the number of provided comments and the mean comment length were generally much larger than their corresponding mean values in each course (see Table 2), it was likely that negative binomial regression would be the better choice. Indeed, negative binomial models had better model fit values (Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood).

Another consideration was the relative frequency of zeroes in the dependent variable. When there are significantly more zeroes than the rest of the distribution would suggest they should occur, adjustments to the modeling approach are needed. As shown in Fig. 3, there were more zeroes than would be expected from the rest of the distribution, especially for the astronomy and psychology courses. Particularly when it makes theoretical sense that there are two separate processes in play, one producing the zeroes and another producing the non-zero quantities, then zero-inflated negative binomial (ZINB) regression is recommended. In the current context, it is plausible that students might be deciding whether to do any reviewing before beginning the reviewing task and then later deciding how many comments to provide if and when they are doing reviews. Based on these considerations and examination of model fit statistics, ZINB was selected as the primary modeling approach. This approach produces two settings of regression weights for each predictor: one for the relative relationship to zero comments and another weight for the close connection to the number of comments when they are greater than zero. Pseudo-R^2 is reported to show model quality separately for predicting zeroes (a logistic regression) and predicting the quantity in the non-zeroes (the negative binomial component). Considering that zero comment length is a missing comment (part of # of comments) rather than a comment whose length is zero, simple negative binomial regression was used for mean comment length, and zeroes were treated as missing (i.e., dropped from the mean calculation).

Finally, to verify the linearity of continuous variables' effects and visualize the relative strength of effects across courses, we also created bins representing roughly a third of each continuous variable's distribution (low, medium, and high) and then plotted the marginal average for that bin.

Results

The Pearson correlations among the predictors for each course are shown in Table 3. None of the core predictors were more than moderately correlated with one another (see upper boxes for each course in Table 3), suggesting no multicollinearity problems. As expected, an examination of the variance inflation factors (VIF) also suggested that multicollinearity was not a significant problem.

Tables 4 and 5 present the results of the regression models for the three courses. Since several of the amount variables (length and number) typically involved much larger numbers than many of the other variables, it is to be expected that they had smaller regression coefficients (i.e., the incremental impact of just one additional comment in an assignment or one additional word within a comment is expected to be small). The difference between the two tables is the choice of dependent variables: number of provided comments (Table 4) vs. length of provided comments (Table 5). The coefficients for zero inflation are presented with a reversed sign to make it easier to see positive effects across the two outcomes. Thus, the inflation coefficient now corresponds to the likelihood of providing at least one comment. In order to better understand effect sizes, the incidence ratio rates (corresponding to the exponent of the beta coefficient) for each core predictor are presented at the bottom. For example, in the psychology and biology courses, having a one-point higher helpfulness rating in the prior assignment more than doubled the likelihood of making at least one comment. A number of comment effects are transformed to be the change with each additional ten comments in one assignment, and the length of comments is transformed to be the change with each additional ten words in one comment.

The number of comments provided on the prior assignment was a good baseline predictor of whether any comments were submitted and the number of comments submitted in the psychology course but not in the other two courses. There may be an effect of the details in the papers to be reviewed (i.e., how many detected problems there are to discuss varying substantially across assignments). By contrast, in all courses, the length of provided comments is a good baseline predictor, reflecting a general individual difference in commenting depth.

Related to the impact of receiving more feedback (either norm-setting or self-efficacy), receiving more comments in the prior assignment was significantly related to providing fewer comments in the next assignment (see Table 4), supporting a self-efficacy effect. Receiving longer comments was related to producing longer comments in two courses (see Table 5; only one effect was significant), providing partial support for norm-setting.

Related to recognition, students who obtained higher helpfulness evaluations in the prior assignment's reviews were consistently more likely to provide at least some comments to their peers (see Table 4) and to provide longer comments (see Table 5). The effect of helpfulness ratings from the prior assignment is visualized in Fig. 4. Note that these are estimated means after controlling for the effects of other variables, including in the regression. Failing to submit any comments was relatively rare in the biology course, so the effect was small in that class.

Table 3 In each course, Pearson intercorrelations among predictors and correlations of predictors with outcome variables (within each box)

	$\#Provide_{J-1}$	$Length\text{-}Provided_{J-1}$	$\#Received_{J-1}$	$Length\text{-}Received_{J-1}$	$Recognition_{J-1}$	$Low\text{-}Score_{J-1}$	$Z\text{-}Score_J$
Psychology							
$Length\text{-}Provided_{J-1}$	−0.14***						
$\#Received_{J-1}$	0.04	−0.05*					
$Length\text{-}Received_{J-1}$	−0.06*	0.06*	−0.28***				
$Recognition_{J-1}$	0.10***	0.26***	−0.04	−0.04			
$Low\text{-}Score_{J-1}$	−0.11***	−0.11***	0.11***	0.19***	−0.07**		
$Z\text{-}Score_J$	0.16***	0.17***	−0.05*	−0.07**	0.08***	−0.22***	
$\#Provided_J$	0.70***	−0.07**	−0.02	−0.02	0.10***	−0.10***	0.13***
$Length\text{-}Provided_J$	−0.08***	0.75***	0.01	0.09***	0.25***	−0.06*	0.17***
Astronomy							
$Length\text{-}Provided_{J-1}$	−0.16***						
$\#Received_{J-1}$	0.37***	−0.16***					
$Length\text{-}Received_{J-1}$	−0.26***	0.12*	−0.25***				
$Recognition_{J-1}$	0.06	0.10*	−0.01	−0.08			
$Low\text{-}Score_{J-1}$	−0.17***	−0.14**	0.03	0.04	−0.04		
$Z\text{-}Score_J$	0.11*	0.18***	0.03	−0.03	0.004	−0.27***	
$Provided_J$	0.10*	0.11*	0.05	−0.16**	0.23***	−0.13**	0.22***
$Length\text{-}Provided_J$	−0.10*	0.68***	−0.14*	0.21***	−0.04	−0.10*	0.18***
Biology							
$Length\text{-}Provided_{J-1}$	−0.02						
$\#Received_{J-1}$	0.46***	0.03					
$Length\text{-}Received_{J-1}$	0.08***	0.17***	0.04				
$Recognition_{J-1}$	0.14***	0.34***	0.02	−0.07**			
$Low\text{-}Score_{J-1}$	0.02	−0.18***	0.07**	0.15***	−0.14***		
$Z\text{-}Score_J$	0.07**	0.18***	−0.01	−0.05*	0.13***	−0.18***	
$\#Provided_J$	−0.03	0.002	−0.08***	−0.06*	0.06**	−0.03	0.02
$Length\text{-}Provided_J$	0.13***	0.64***	0.04*	0.13***	0.23***	−0.14***	0.26***

Note: ***=p < .001, **=p < .01, *p < .05

Related to mastery experiences, students who did well on the current assignment consistently were more likely to provide at least some comments to their peers (see Table 4) and to provide longer comments (see Table 5). Further, in the astronomy class, doing well on the current assignment also predicted the number of comments given, in addition to submitting at least some comments. Interestingly, there appears to be no additional effect of task score from the prior assignment ($Low\text{-}Score_{J-1}$) on the number of comments provided or the length of the comments provided. Figure 5 visualizes the effects of the score on the current assignment on provided length, and Fig. 6 visualizes the effects of the prior helpfulness of provided length. Note that the effects were always monotonic and very close to linear, supporting the use of linear regression models.

For the control variables (see Tables 4 and 5), the course offerings within a discipline sometimes differed in the number of comments, with one of the courses producing more comments than the other. Additionally, sometimes commenting behaviors change across assignments, but in relatively variable ways. In the astronomy course, students increased the number of comments provided across assignments but then gave shorter comments. In psychology, students just gave shorter comments in later assignments. In biology, students gave fewer but longer comments in later assignments.

General discussion

The goal of this study was to uncover experiential factors that could explain variation in student levels of participation in providing peer feedback comments so that future interventions could be designed to ensure more equitable experiences for peer feedback receivers. This study tested three hypotheses involving the relationships of different experiential aspects during peer feedback cycles to changes in whether feedback was provided and how much feedback students provided to their peers (number of comments and length of comments). The study used a strong context variation strategy (different kinds of courses in different disciplines in different universities) to uncover patterns that are, therefore, more likely to be generally observed across contexts. Table 6 summarises the results of the hypothesis-testing results.

Evaluating each tested hypothesis

There was weak support for norm-setting: just one positive result for comment length. The relationship to the number of comments was significantly negative in all three courses, supporting the alternative hypothesis related to self-efficacy effects. A lack of a significant relationship, particularly in a single study, must be interpreted with caution: perhaps the effect is small, and the power of the study was not large enough to detect such a small effect. However, the *N* within two of the three contexts was quite substantial, suggesting that any effect would not be practically meaningful even if eventually statistically significant. A possible interpretation is that norm-setting likely does still occur but at a different temporal scale. For example, perhaps the norms are set early on (Turpen & Finkelstein, 2010) and then not substantially changed with additional experience in a given context. Similarly, norms may be set at a slower time scale (e.g., based on the amount of feedback received in the last three or four assignments, not just the last assignment).

Table 4 For the zero-inflated negative binomial (ZINB) regression models predicting the number of provided comments (Count) and the presence of any provided comments (Not Zero), the estimated coefficients for each predictor, the overall model fit statistics, and the effect sizes for core predictors (incidence ratio rate) within each course

Predictor	Psychology		Astronomy		Biology	
	Count	Not Zero	Count	Not Zero	Count	Not Zero
Baseline						
# Provided$_{J-1}$	0.03***	0.06*	-0.01***	-0.04	0.00	-0.02
Core predictors						
#Received$_{J-1}$	-0.004***	-0.003	-0.01***	0.04	-0.003*	-0.02
Recognition$_{J-1}$	0.01	0.77*	0.05***	0.05*	0.02	0.77*
Low-Score$_{J-1}$	-0.003	-0.42	-0.03	-0.79	0.003	0.04
Z-Score$_{J}$	-0.023	0.24*	0.04**	0.37*	0.005	0.49*
Control variables						
Course	-0.05*	0.38	NA	NA	-0.12***	-2.9**
J	0.08	-0.15	0.34***	-0.14	-0.10***	0.05
N	1,506		373		1,849	
Pseudo R^2	0.09	0.09	0.10	0.09	0.06	0.17
Effect sizes (incidence ratio rate)						
#Received$_{J-1}$*/10*	0.96	0.97	0.90	1.49	0.97	0.82
Recognition$_{J-1}$	1.01	2.16	1.05	1.05	1.02	2.16
Low-Score$_{J-1}$	1.00	0.66	0.97	0.45	1.00	1.04
Z-Score$_{J}$	0.98	1.27	1.04	1.45	1.01	1.63

Note: ***=$p<.001$, **=$p<.01$, *=$p<.05$, NA=Not applicable

Table 5 For the negative binomial regression models predicting the length of provided comments, the estimated coefficients and fit statistics within each course

Predictor	Psychology		Astronomy		Biology	
	Coef.	Sig.	Coef.	Sig.	Coef.	Sig.
Baseline						
Length-Provided$_{J-1}$	0.002	***	0.0002	***	0.0001	***
Core predictors						
Length-Received$_{J-1}$	0.0002	*	0.003	^{t}	-0.00	
Recognition$_{J-1}$	0.10	***	-0.04	*	0.04	*
Low-Score$_{J-1}$	0.005		0.001		-0.02	
Z-Score$_{J}$	0.02	***	0.04	**	0.06	***
Control variables						
Course	0.008		NA	NA	0.02	
J	-0.04	***	-0.10	***	0.05	***
N	1,472		349		1,835	
Pseudo R^2	0.09		0.10		0.06	
Effect sizes (incidence ratio rate)						
Length-Received$_{J-1}$*/10*	1.01		1.03		1.00	
Recognition$_{J-1}$	1.11		0.96		1.04	
Low-Score$_{J-1}$	1.01		1.00		0.98	
Z-Score$_{J}$	1.02		1.04		1.06	

Note: ***=$p<.001$,**=$p<.01$, *=$p<.05$, ^{t}=p<.1, NA=Not applicable

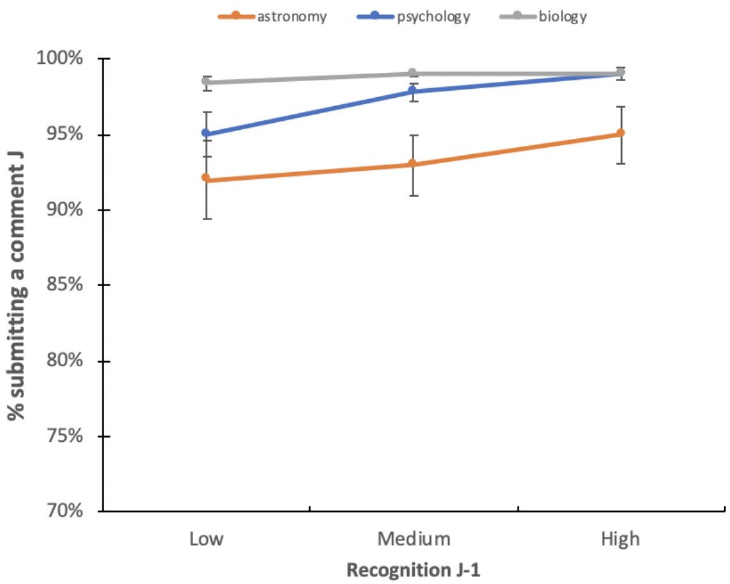

Fig. 4 Within each course, the percentage of students submitting any comments on the J^{th} assignment (with SE bars) as a function of having received a low, medium, or high amount of recognition for providing helpful reviews on the $(J\text{-}1)^{th}$ assignment, controlling for other predictors

There was strong support for reinforcement: in all three courses, being recognized for good reviewing predicted growth in the number of comments provided, and there was also a significant growth in comment length in two of the three courses. This finding suggests the prior weak support obtained by Zong et al. (2022) may have been due to the rare and binary nature of recognition for good reviewing that was found in that system, in contrast to the common and more nuanced recognition for good reviewing found in systems like *Peerceptiv, ELI Peer Review, and Kritik*. However, there might have been other differences at play: cultural contexts or programming vs. writing. Future research will have to examine possible explanations for variations in effect sizes.

The relationship of predictors to outcomes for the self-efficacy-related pathways was mixed. Students submitted shorter comments and were less likely to submit any comments when they performed poorly on the current assignment. Although self-efficacy was not directly measured, this effect was expected as a kind of self-efficacy/personnel regulation effect. Self-efficacy has been raised in interviews and survey studies of peer assessment (Liu & Carless, 2006; Kaufman & Schunn, 2011). The current study adds behavioral indicators to studies of this effect and clarifies that the results seem to be most relevant to deciding whether to provide any feedback, in contrast to the location of the recognition effect.

Performance evaluations on the prior task did not relate to the number of comments or comment length in any of the three courses. The lack of this effect was interesting, while there was an effect for current assignment evaluations. It may be that confidence varies substantially from assignment to assignment based on various possible factors (e.g., different content or skills assessed by the assignment, level of investment in the assignment), and that confidence-informing experience in the current assignment is what matters most. Alterna-

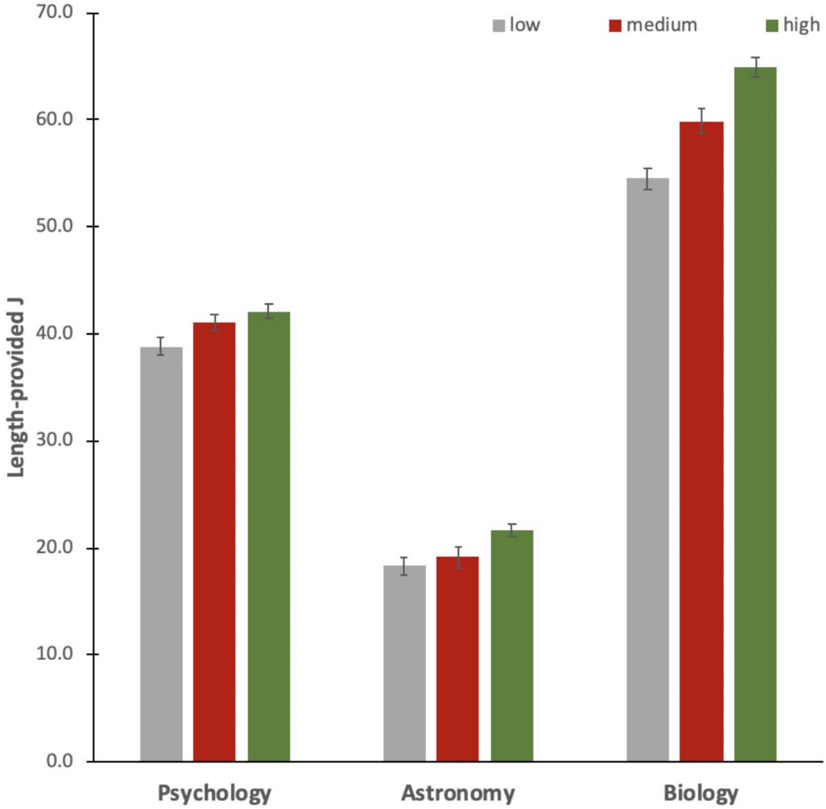

Fig. 5 Within each course, the number of words per comment provided on the J^{th} assignment (with SE bars) as a function of whether the student's contribution on the same assignment received a low, medium, or high score, controlling for other predictors

tively, it could be that the comments mattered more than the ratings in the prior assignment to shape student confidence. This would explain why the number of comments but not the ratings were significant negative predictors.

Contextual variation

Three relationships of recent experience with providing feedback were observed only within one of the three contexts: receiving more words on the prior round (*Length-Received*$_{J-1}$) predicted giving more words (*Length-Provided*$_J$) only in the psychology courses (see Table 5); having a high self-evaluation score on the current assignment indicated giving more occasional comments only in the astronomy courses (see Table 4); being recognized for providing high-quality comments was associated with being more likely to provide more comments in astronomy (see Table 4). One of the three significant relationships was only significant at the $p < .05$ level, which might be a false positive. However, it was in the expected direction. Another of the three involved a relationship that could have meaningfully been positive or negative based on different underlying mechanisms (norm-setting or

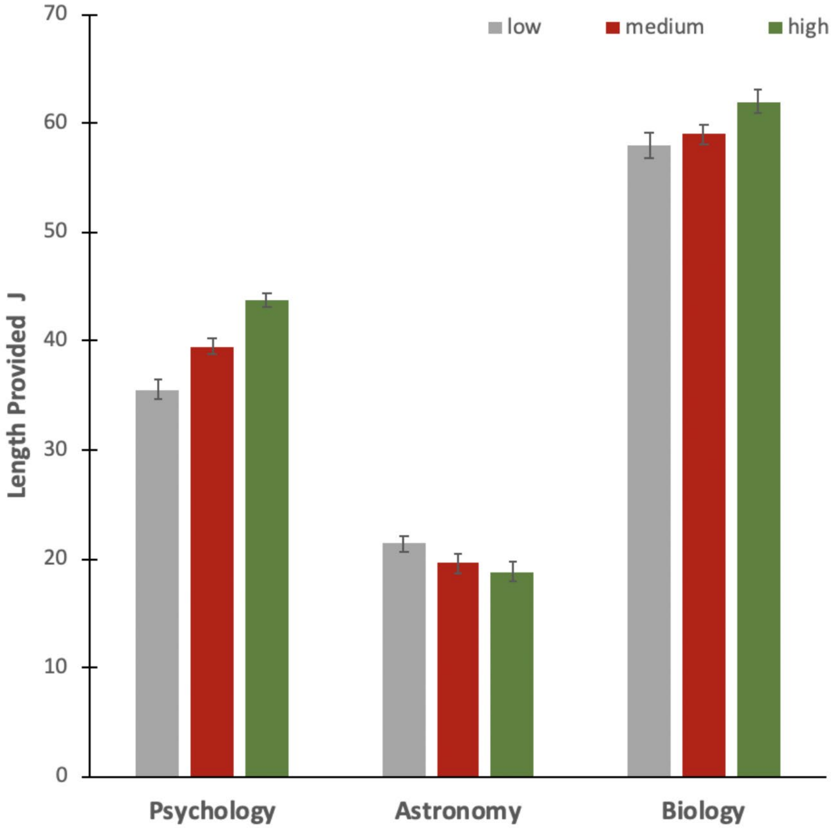

Fig. 6 Within each course, the mean number of words per comment provided to peers on the J^{th} assignment (with SE bars) as a function of whether the student's comments provided on the prior assignment received recognition as being of low, medium, or high helpfulness from the comment recipients, controlling for other predictors

self-efficacy influencing), and this effect was significant at $p < .001$. This effect, therefore, was unlikely to be a false positive. Further, the estimated coefficients in the non-significant cases were quite different from the estimated coefficients in the significant cases (and outside each other's 95% confidence intervals). Thus, there appears to be a meaningful variation of effects by context.

Assuming the relationships varied by context are indeed types of relationships that do regularly depend upon context factors, it now becomes a new research question to understand what the moderators are and why they occur. Theoretically, varying amounts of structure and support comment contents might play an important role. For example, if there is clear guidance on what to include within a comment, the length of the comment may vary less across reviews. Similarly, if there is clear guidance on which issues to discuss or the focus is on a smaller set of issues, the number of comments may vary less across reviews. The selected courses did vary substantially in the amount and kinds of details included in the peer feedback rubrics.

Table 6 Summary of hypotheses tests in three courses

Hypotheses	Results of hypothesis tests in three courses	
	# of comments	Length of comments
H1 (self-efficacy): High task quality in the prior assignment will predict more active participation in peer review.	3–	3–
H2 (self-efficacy): More feedback (which is mostly negative) in the prior assignment predicts less active participation in peer review	3+	3–
H3 (self-efficacy): Higher task quality in the current assignment will predict more active participation in peer review.	3+	3+
H4 (norm-setting): Greater amounts of feedback received in the prior assignment predicts more active participation in peer review	3–	1 + 2–
H5 (reinforcement): Higher helpfulness ratings from peers in the prior assignment will predict more active participation in peer review.	3+	2 + 1–

Note: "+" = supported "–" = not supported

Other explanations of contextual variation are also possible, particularly ones that use the type of course (general education with a broad set of student backgrounds vs. majors-only courses with a narrower set of backgrounds) or the discipline or writing task. In the more objective natural sciences disciplines, feedback on accuracy could have stronger effects on student self-efficacy (i.e., there is less room for disagreement). Alternatively, the breadth of student backgrounds in the course (e.g., the substantial number of non-majors in the psychology and astronomy courses) may also increase the likelihood that students, especially those with weaker backgrounds in the course discipline, will have self-efficacy concerns.

Practical implications

Peer reviewing has proven to assess student performance effectively and provide a robust platform for students to learn from one another. Nevertheless, the willingness to participate at all (Liu & Carless, 2006; Huisman et al., 2018) or with substantial levels of feedback (Cho & MacArthur, 2011; Tsivitanidou et al., 2018) has proven to be a common bottleneck. With online peer assessment, interfaces can be adapted to increase the amount of feedback that students give (e.g., through minimum word counts in *CPR* or the minimum number of comments in *Peerceptiv*), but still, other strategies to improve participation are needed.

While not conclusive, the current research provides some suggestions for directions that are more likely to be productive. In particular, it is unlikely that simple exposure and experience with peer feedback systems will gradually improve student participation. Allowing for the possibility of rewards for reviewing can increase the quality of comments/length of comments (Patchan et al., 2018; Zong et al., 2021b). Using systems like *Peerceptiv* or *Kritik*, which directly recognize review quality, involves an easy-to-implement instructor action. A related strategy might encourage students to express gratitude more commonly for the feedback received, as opposed to only complaining about problems in the feedback they received. Alternatively, automated systems could be used to give students praise for producing longer, more in-depth feedback (Ramachandran et al., 2017).

However, the recognition effects were relatively small; only the effect on choosing to submit any comments at all was large. Instead, the current study suggests that student self-efficacy is a crucial factor to target, with larger and more consistent effects on comment length and more consistent effects on whether to submit any comments at all. Thus, when students are lagging in producing any reviews or just a small number of reviews, instructors might intervene with comments about the effectiveness of the comments they did produce in this round or prior rounds. Alternatively, the instructor could provide some training on what makes for effective feedback (Berg, 1999; Sluijsmans et al., 2002; Min, 2016).

Other possible strategies might involve avoiding the overall effect of fatigue from too many peer assessment assignments alongside a greater workload across courses later in the semester, as students in these courses either reduced the number of comments they provided or reduced the length of the comments they provided in later assignments. Instructors might reduce the number of reviews that are assigned or focus the reviews on smaller sections of a submitted assignment. Another strategy might involve small reductions in assignments' difficulty to reduce students' self-efficacy concerns for these assignments, thereby increasing students' feedback.

Limitations and Future Research

The current study was fundamentally correlational in examining the role of experience in shaping the amount of feedback provided and its investigation of moderation of findings by course context. Therefore, strong causal claims cannot be made based on the current data. However, as an initial exploration of a novel research topic, the study has provided evidence in an externally-valid way (with real interfaces on real classroom assignments) of potentially important empirical phenomena. Further, by examining change over time with various statistical controls, the reverse causality and obvious third-variable confounds have been ruled out (e.g., general differences across reviewers in the amount of feedback provided), at least for the role of experience factors. Future studies should be conducted that more carefully control student experiences (e.g., by experimentally manipulating recognition or relative performance feedback) to directly test the causal status of these factors.

Another open question relates to the underlying causes. We have posited that self-efficacy could explain the observed patterns in the current data, but self-efficacy was not directly measured. This is particularly important to understand the effects of receiving more comments. It would be challenging to measure self-efficacy across many assignments in an actual classroom regularly, but the role of feedback on self-efficacy could be tested on a smaller number of assignments or within an experimental study. The current findings justify the investment in such future empirical work.

A third open question relates to other dimensions of feedback quality. Here we have focused on the amount of feedback provided in terms of the number of comments. Comments can vary substantially in terms of their depth (e.g., with or without explicitly identifying the problem, explaining the situation, providing a constructive suggestion, and explaining the constructive suggestion; Wu and Schunn, 2020). These comment elements also shape the learning opportunities for feedback providers and receivers (Cho & Schunn, 2007; Wu & Schunn, 2023). Feedback experience could shape whether and how much reviewers include these elements in their comments as well. Pragmatically, automated systems could be developed to automatically classify these aspects of peer feedback (Ramachandran et al., 2017) to make it pragmatically possible to do such research on the scale of large classes with multiple peer feedback assignments.

Authors' contributions Not applicable (for blinded review).

Funding This research study was funded by the Humanities and Social Sciences project, Ministry of Education in China (21YJA880065).

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All co-authors consent to the publication of this study.

Conflicts of interest/Competing interests Not applicable.

References

Abubakar, A. M., Elrehail, H., Alatailat, M. A., & Elçi, A. (2019). Knowledge management, decision-making style and organizational performance. *Journal of Innovation & Knowledge*, *4*(2), 104–114. https://doi.org/10.1016/j.jik.2017.07.003

Adachi, C., Tai, J. H. M., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education*, *43*(2), 294–306. https://doi.org/10.1080/02602938.2017.1339775

Applebee, A., & Langer, J. (2011). *The national study of writing instruction: Methods and procedures (CELA Report)*. Retrieved from University of Albany, NY: Center on English Learning and Achievement website. http://www.albany.edu/cela/reports/NSWI_2011_methods_procedures.pdf

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education*, *27*(5), 427–441. https://doi.org/10.1080/0260293022000009302

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Berg, E. C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, *8*(3), 215–241. https://doi.org/10.1016/S1060-3743(99)80115-5

Brown, G. T., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology*, *86*(4), 606–629. https://doi.org/10.1111/bjep.12126

Chang, C. Y. H. (2016). Two decades of research in L2 peer review. *Journal of Writing Research*, *8*(1), 81–117. https://doi.org/10.17239/jowr-2016.08.01.03

Chapman, O. L., & Fiore, M. A. (2000). Calibrated peer review™. *Journal of Interactive Instruction Development*, *12*(3), 11–15.

Chen, C. H., & Chiu, C. H. (2016). Collaboration scripts for enhancing metacognitive self-regulation and mathematics literacy. *International Journal of Science and Mathematics Education*, *14*(2), 263–280. https://doi.org/10.1007/s10763-015-9681-y

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education*, *24*(3), 301–314. https://doi.org/10.1080/0260293990240304

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 and Education*, *48*(3), 409–426. https://doi.org/10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Wilson, R. (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

Crinon, J. (2012). The dynamics of writing and peer review at primary school. *Journal of Writing Research*, *4*(2), 121–154. https://doi.org/10.17239/jowr-2012.04.02.2

Daniels, A. C. (2016). *Bringing out the best in people: How to apply the astonishing power of positive reinforcement*. New York, NY: McGraw Hill Professional.

Deiglmayr, A. (2018). Instructional scaffolds for learning from formative peer assessment: Effects of core task, peer feedback, and dialogue. *European Journal of Psychology of Education*, *33*(1), 185–198. https://doi.org/10.1007/s10212-017-0355-8

Dinsmore, D., Alexander, P., & Loughlin, S. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*, 391–409. https://doi.org/10.1007/s10648-008-9083-6

Feng, S., Wong, Y. K., Wong, L. Y., & Hossain, L. (2019). The internet and Facebook usage on academic distraction of college students. *Computers & Education*, *134*, 41–49. https://doi.org/10.1016/j.compedu.2019.02.005

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, *37*(1), 16–55. https://doi.org/10.2307/357381

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404. https://doi.org/10.1037/0033-2909.118.3.392

Grogger, J. T., & Carson, R. T. (1991). Models for truncated counts. *Journal of Applied Econometrics*, *6*(3), 225–238. https://doi.org/10.1002/jae.3950060302

Haaga, D. A. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology*, *20*(1), 28–32. https://doi.org/10.1207/s15328023top2001_5

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education*, *40*(1), 151–164. https://doi.org/10.1080/02602938.2014.893418

Harris, L. R., Brown, G. T., & Harnett, J. A. (2015). Analysis of New Zealand primary and secondary student peer-and self-assessment comments: Applying Hattie and Timperley's feedback model. *Assessment in Education: Principles Policy & Practice*, *22*(2), 265–281. https://doi.org/10.1080/0969594X.2014.976541

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

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Kankanhalli, A., Tan, B. C., & Wei, K. K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS Quarterly*, *29*(1), 113–143. https://doi.org/10.2307/25148670

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science*, *39*(3), 387–406. https://doi.org/10.1007/s11251-010-9133-6

Li, H., Xiong, Y., Zang, X., Kornhaber, M., Lyu, Y., Chung, K. S., & 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

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

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

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

Mangelsdorf, K. (1992). Peer reviews in the ESL composition classrooms: What do the students think? *ELT Journal*, *46*(3), 274–284. https://doi.org/10.1016/j.jslw.2008.06.002

Marcoulides, G. A., & Simkin, M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business*, *70*(4), 220–223. https://doi.org/10.1080/08832323.1995.10117753

Margolis, H., & McCabe, P. P. (2003). Self-efficacy: A key to improving the motivation of struggling learners. *Preventing School Failure: Alternative Education for Children and Youth*, *47*(4), 162–169. https://doi.org/10.1080/10459880309603362

Martin, A. J., & Evans, P. (2018). Load reduction instruction: Exploring a framework that assesses explicit instruction through to independent learning. *Teaching and Teacher Education*, *73*, 203–214. https://doi.org/10.1016/j.tate.2018.03.018

McLeod, M., Hart-Davidson, W., & Grabill, J. (2013). Theorizing & Building Online writing environments: User-centered design beyond the interface. In G. Pullman, & B. Gu (Eds.), *Designing web-based applications for 21st Century writing classrooms*. Amityville, NY: Baywood Press.

Meusen-Beekman, K. D., Joosten-ten Brinke, D., & Boshuizen, H. P. (2016). Effects of formative assessments to develop self-regulation among sixth grade students: Results from a randomized controlled intervention. *Studies in Educational Evaluation*, *51*, 126–136. https://doi.org/10.1016/j.stueduc.2016.10.008

Min, H. T. (2005). Training students to become successful peer reviewers. *System*, *33*(2), 293–308. https://doi.org/10.1016/j.system.2004.11.003

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. https://doi.org/10.1016/j.jslw.2006.01.003

Min, H. T. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing*, *31*, 43–57. https://doi.org/10.1016/j.jslw.2016.01.004

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research*, *2*(3), 192–222. https://doi.org/10.1287/isre.2.3.192

Moore, C., & Teather, S. (2013). Engaging students in peer review: Feedback as learning. *Issues in Educational Research*, *23*(Suppl.), 196–211. https://doi.org/10.3316/informit.354576626678153

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing: A case study from geography. *Innovations in Education and Training International*, *32*(4), 324–335. https://doi.org/10.1080/1355800950320404

Paltridge, B. (2015). Referees' comments on submissions to peer-reviewed journals: When is a suggestion not a suggestion? *Studies in Higher Education*, *40*(1), 106–122. https://doi.org/10.1080/03075079.2013.818641

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

Patchan, M. M., Schunn, C. D., & 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

Popp, J. S., & Goldman, S. R. (2016). Knowledge building in teacher professional learning communities: Focus of meeting matters. *Teaching and Teacher Education*, *59*, 347–359. https://doi.org/10.1016/j.tate.2016.06.007

Ramachandran, L., Gehringer, E. F., & Yadav, R. K. (2017). Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, *27*(3), 534–581. https://doi.org/10.1007/s40593-016-0132-x

Rodriguez, C., Hudson, R., & Niblock, C. (2018). Collaborative learning in architectural education: Benefits of combining conventional studio, virtual design studio and live projects. *British Journal of Educational Technology*, *49*(3), 337–353. https://doi.org/10.1111/bjet.12535

Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, *33*(1), 75–99. https://doi.org/10.1007/s10212-017-0339-8

Schunn, C. D., Godley, A. J., & 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

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, *19*(1), 69–75. https://doi.org/10.1080/03075079412331382153

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

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.1080/02602938.2018.1548559

Tsivitanidou, O. E., Constantinou, C. P., Labudde, P., Rönnebeck, S., & Ropohl, M. (2018). Reciprocal peer assessment as a learning tool for secondary school students in modeling-based learning. *European Journal of Psychology of Education*, *33*(1), 51–73. https://doi.org/10.1007/s10212-017-0341-1

Turpen, C., & Finkelstein, N. D. (2010). The construction of different classroom norms during peer instruction: Students perceive differences. *Physical Review Special Topics-Physics Education Research*, *6*(2), 020123. https://doi.org/10.1103/PhysRevSTPER.6.020123

Valero Haro, A., Noroozi, O., Biemans, H. J., & Mulder, M. (2019). The effects of an online learning environment with worked examples and peer feedback on students' argumentative essay writing and domain-specific knowledge acquisition in the field of biotechnology. *Journal of Biological Education*, *53*(4), 390–398. https://doi.org/10.1080/00219266.2018.1472132

van Blankenstein, F. M., Truțescu, G. O., van der Rijst, R. M., & Saab, N. (2019). Immediate and delayed effects of a modeling example on the application of principles of good feedback practice: A quasi-experimental study. *Instructional Science*, *47*(3), 299–318. https://doi.org/10.1007/s11251-019-09482-5

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

VanStelle, S. E., Vicars, S. M., Harr, V., Miguel, C. F., Koerber, J. L., Kazbour, R., & Austin, J. (2012). The publication history of the Journal of Organizational Behavior Management: An objective review and analysis: 1998–2009. *Journal of Organizational Behavior Management*, *32*(2), 93–123. https://doi.org/10.1080/01608061.2012.675864

Wang, Y., Li, H., Feng, Y., Jiang, Y., & Liu, Y. (2012). Assessment of programming language learning based on peer code review model: Implementation and experience report. *Computers & Education*, *59*(2), 412–422. https://doi.org/10.1016/j.compedu.2012.01.007

Wang, X. M., Hwang, G. J., Liang, Z. Y., & Wang, H. Y. (2017). Enhancing students' computer programming performances, critical thinking awareness and attitudes towards programming: An online peer-assessment attempt. *Journal of Educational Technology & Society*, *20*(4), 58–68. https://www.jstor.org/stable/26229205

Whicher, A., Harris, C., Beverley, K., & Swiatek, P. (2018). Design for circular economy: Developing an action plan for Scotland. *Journal of Cleaner Production*, *172*, 3237–3248. https://doi.org/10.1016/j.jclepro.2017.11.009

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

Wu, Y., & Schunn, C. D. (2021). 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*, 102160. https://doi.org/10.1016/j.cedpsych.2023.102160

Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. *Journal of Research in Science Teaching*, *45*(9), 1036–1058. https://doi.org/10.1002/tea.20195

Zhang, F., Schunn, C. D., & Baikadi, A. (2017). Charting the routes to revision: An interplay of writing goals, peer comments, and self-reflections from peer review. *Instructional Science*, *45*(5), 679–707. https://doi.org/10.1007/s11251-017-9420-6

Zimmerman, B. J., & Risemberg, R. (1997). Caveats and recommendations about self-regulation of writing: A social cognitive rejoinder. *Contemporary Educational Psychology*, *22*(1), 115–122. https://doi.org/10.1006/ceps.1997.0921

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

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

Zong, Z., Schunn, C. D., & Wang, Y. (2022). What makes students contribute more peer feedback? The role of within-course experience with peer feedback. *Assessment and Evaluation in Higher Education*, *47*(6), 972–983. https://doi.org/10.1080/02602938.2021.1968792

Zou, Y., Schunn, C. D., Wang, Y., & Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assessment and Evaluation in Higher Education*, *43*(5), 800–811. https://doi.org/10.1080/02602938.2017.1409872