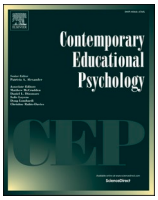
Contemporary Educational Psychology 74 (2023) 102210

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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



Self-regulation of peer feedback quality aspects through different dimensions of experience within prior peer feedback assignments

Yi Zhang [a,*], Christian D. Schunn [b]

[a] College of Education for the Future, Beijing Normal University, No. 18 Jingfeng Road, Beijing Normal University, Zhuhai 519087, Guangdong Province, China
[b] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Avenue, Pittsburgh, PA 15260, USA

ARTICLE INFO

Keywords:
Peer review quality
Self-regulated learning
Peer feedback literacy

ABSTRACT

Although peer review is a widely-used pedagogical technique, its value depends upon the quality of the reviews that students produce, and much research remains to be done to systematically study the nature, causes, and consequences of variation in peer review quality. We propose a new framework that conceptualizes five larger dimensions of peer review quality and then present a study that investigated three specific peer review quality constructs in a large dataset and further explored how these constructs change through different types of self-regulation peer reviewing experiences. Peer review data across multiple assignments were analyzed from 2,092 undergraduate students enrolled in one of three offerings of a biology course at a large public research university in the United States. Peer review quality was measured in terms of comment amount, comment accuracy, and rating accuracy; the measures of reviewing experience focused upon self-regulated learning factors such as practice, feedback, others' modeling, and relative performance. Meta-correlation (for testing reliability, separability, and stability) and *meta*-regression (as a time-series analysis for testing the relationship of changes across assignments in reviewing quality with experiences as reviewer and reviewee) are used to establish the robustness of effects and meaningful variation of effects across course offerings and assignments. Results showed that there were three meaningful review quality constructs (i.e., were measured reliably, separable, and semi-stable over time). Further, all three showed changes in response to previous reviewer and reviewee experiences, but only feedback helpfulness, in particular, showed effects of all four examined types of self-regluation experiences (practice, feedback, others' modeling, and relative performance). The findings suggest that instructors can improve review quality by providing comment prompt scaffolds that lead to longer comments as well as by matching authors with similarly performing reviewers.

1. Introduction

Peer review is a complex kind of learner-focused educational activity with high levels of learner self-regulation that combines providing and receiving comments (often called peer feedback) and ratings (often called peer assessment) based on documents from peers in their class (Li et al., 2016; Sanchez et al., 2017). It is becoming widely used in education, especially with the support of online tools (Dmoshinskaia et al., 2021a; Zheng et al., 2020). Online peer review offers a number of supportive features, such as anonymous reviewing (Panadero & Alqassab, 2019), training/calibration processes (Balfour, 2013; Suen, 2014), scaffolds for ratings and comments (Cho & Schunn, 2007; Latifi et al., 2021; Zheng et al., 2020), and accountability for higher quality reviewing (Misiejuk & Wasson, 2021). In addition, online peer review systems automatically collect extensive data on peer review learner behaviors, creating new opportunities to study the complex aspects of peer review behaviors and underlying competencies.

Research on peer review in education has greatly expanded in recent years (Double et al., 2020), and there have been many recently published systematic reviews (Misiejuk & Wasson, 2021; Panadero & Alqassab, 2019; van Popta et al., 2017) and *meta*-analyses (Chang et al., 2021; Li et al., 2016; Li et al., 2020; Li et al., 2021; Sanchez et al., 2017; Double et al., 2020; Huisman et al., 2019; Thirakunkovit & Chamcharatsri, 2019; Zheng et al., 2020) focused on different aspects of peer review or its impacts for different learner populations. Overall, the *meta*-analyses indicate that peer assessments (i.e., the ratings) are typically reliable and valid and that peer feedback (i.e., the providing and receiving comments) has moderate-sized effects on student writing,

* Corresponding author.
E-mail addresses: yizhan_g@bnu.edu.cn (Y. Zhang), schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.cedpsych.2023.102210

Available online 24 July 2023

understanding, and attitudes. These benefits have been observed in K-12 and higher education, native and foreign language classrooms, and with and without technology support. Learning through peer review has been theoretically grounded in terms of focused practice with feedback as a kind of deliberate practice (Kellog & Whiteford, 2009; Wu & Schunn, 2021a), social interaction with similarly skilled individuals as a zone of proximal development (Havnes, 2008; Shabani et al., 2010), and learning by observing peer comments and reflecting on them as a kind of self-regulated learning (Alemdag, & Yildirim, 2022; Zong et al., 2021b). And in research on writing, the students' SRL strategies has been found to be significantly correlated with their writing performance (Nückles et al., 2020; Teng, & Zhang, 2018). Here we focus on the theoretical framework of self-regulated learning to understand changes in reviewing behavior over time, where producing comments while reviewing can be considered a kind of writing task.

Similar to other work examining the character and quality of self-regulated learning, a number of studies have found that the quality of peer assessments and peer feedback is not always satisfactory (Misiejuk & Wasson, 2021; Xiong & Schunn, 2021; Yuan et al., 2016). Lower-quality assessments might arise from a lack of accountability (Patchan et al., 2018) or a lack of understanding of criteria (Könings et al., 2019). Lower-quality feedback might arise from an unwillingness to share negative comments because of concerns about harming friendships (Kilickaya, 2017) or as a reaction to having received negative evaluations (Lin et al., 2001) or lower-quality feedback (Zong et al., 2021b, 2022a) in a prior assignment. Importantly, instructors may be less likely to use peer review if the assessment quality is low (Kilickaya, 2017). Just as importantly, students are less likely to benefit from providing or receiving feedback if the feedback in that context tends to be of low quality (Cambre et al., 2018): students can develop a false sense of confidence (Kilickaya, 2017), they can be less likely to revise their work (Sommers, 1982), they can lower performance standards (Yeager et al., 2014), or they can lead other students also to provide lower quality peer feedback (Zong et al., 2022b).

However, much research remains to be done to systematically study the causes and consequences of variation in the quality of peer assessments and peer feedback to understand better how to support students' self-regulation of peer feedback. We argue that part of the problem is not enough work has been done to systematically conceptualize different general dimensions and specific constructs of peer review quality. We provide a new larger conceptual framework of the main general dimensions for peer review quality and then present a study that empirically examines three specific feedback quality constructs situated within that framework.

There is also a conceptual challenge in the existing research literature related to grain-size. Feedback quality, similar to other kinds of self-regulated learning behaviors, has been conceptualized as a property of an action or moment in time (i.e., the review) and, therefore, jointly influenced by three things: characteristics of the reviewer, the reviewed document, and the context of the review (Xiong & Schunn, 2021). However, particular research can also focus on only one of the three levels, averaging over the other two. For example, studies can focus on the level of the reviewed document (e.g., did this object receive accurate ratings?; did this object receive useful feedback?; Huisman et al., 2019; Kaufman & Schunn, 2011; Schillings et al., 2021) or at the level of the context (e.g., are ratings more reliable under anonymous conditions or with accountability pressures; Patchan et al., 2018). Critically, peer review quality could also be conceptualized as a property of the reviewer, as part of peer review literacy (Dong et al., 2023), which is a subtype of general feedback literacy (Zhan, 2021). That is, the reviewer has some knowledge and skills related to producing peer reviews and some attitudes towards peer reviews (Nieminen & Carless, 2022), and the output of such knowledge, skills, and attitudes will be a tendency to provide higher or lower quality reviews. Little prior work has examined peer review literacy (Dong et al., 2023). Given that peer review is such a learner-centric pedagogy, it is particularly important that students develop capacities to effectively participate in peer review (i.e., as productive self-regulated learners), which in turn means that practitioners and researchers need ways to conceptualize different aspects of peer review literacy better.

We take up that challenge in the current study by focusing on conceptualizations of peer review quality as a reviewer characteristic that is self-regulated through experiences with peer feedback. We test this conceptualization as a reviewer chracteristic by examining the extent to which different specific review quality constructs appear to have some stability at the reviewer level and can be empirically distinguished from one another at the reviewer level. Having established the utility of that conceptualization, we then turn to study the self-regulation processes by examining which experiences as a reviewer and reviewee are positively or negatively associated with changes in peer feedback quality.

1.1. Conceptualizing review quality at the reviewer level: Comment amount, comment accuracy, and rating accuracy

As a new conceptual framework for peer review quality, we argue that the many specific constructs of peer review quality that have been studied can be organized according to five larger dimensions of peer feedback quality, and each of them can be measured at the grain size of a comment/rating, a review, a document, a reviewer, or a context level. For example, as the first dimension in our framework, some researchers have examined feedback quality in terms of *feedback features* (i.e., was the feedback phrased in useful ways?), either in terms of cognitive features that influence feedback understandability or affective features that influence feedback agreement or willingness to implement (e.g., Patchan et al., 2018; Jin et al., 2022; Kerman et al., 2022; Noroozi et al., 2022; Tan & Chen, 2022). Under the second dimension of peer review quality, researchers have examined feedback quality in terms of *feedback content* (i.e., was the feedback about the right topics?), such as alignment to peer assessment rubric, focus on important problems in the document, or attending to more generally important topics (e.g., Gao et al., 2019; Huisman et al., 2019; van den Bos, & Tan, 2019). The third dimension is *feedback impact* (i.e., did the feedback impact the receiver?), such as the amount of document revision by the feedback recipient or level of engagement with the feedback by the feedback recipient (e.g., Strijbos et al., 2010; Vanderhoven et al., 2015; Wu & Schunn, 2020). A fourth dimension is *review accuracy* (i.e., were the feedback comments or ratings accurate?), including the actual or perceived accuracy of descriptions of problems and recommendations for improvements (e.g., Wu & Schunn, 2021b; Van Steendam et al., 2010) and rating reliability or validity (e.g., Cho et al., 2006; Li et al., 2016). Finally, the fifth dimension is *feedback amount* (i.e., how much feedback was provided?), whether in terms of the number of reviews or comments produced or the length of the comments produced (e.g., Jin et al., 2022; Patchan et al., 2018; Zong et al., 2021a).

In terms of the specific grain-size of a review quality construct within any of the five dimensions, while it is possible to produce measures at the fine-grain size of a rating/comment or a review, it is also logically possible to produce aggregate values at larger grain-sizes, such as a score for a reviewer (e.g., taking the average length across all comments produced by a reviewer on an assignment), a document, or a whole assignment. In the current study, we focus on the level of a reviewer. Past researchers that analyzed peer review quality at the reviewer level have tended to focus on only a few of the five broad feedback quality dimensions: feedback amount (e.g., Patchan et al., 2018; Zong et al., 2021a), feedback features (e.g., Jin et al., 2022; Patchan et al., 2018), or feedback accuracy (e.g., Cui et al., 2021). These studies involved training interventions in improving feedback quality (e.g., Cui et al., 2021), intervention studies to see whether reviewing scaffolds or accountability methods improved reviewing quality (Patchan et al., 2018), or regression studies to examine whether experiences in prior courses or assignments were associated with better-reviewing quality (Zong et al., 2021a).

In the present study, we focused on three feedback quality constructs. Two constructs are drawn from the fourth dimension (review accuracy): rating accuracy and comment accuracy constructs. One construct is drawn from the fifth dimension (feedback amount). These three peer review quality constructs are measured at the reviewer level and are studied as an initial test of the conceptual framework because these are more easily assessed in larger datasets and thus of more pragmatic value to the field. By easily assessed, we mean that the specific indicators of review quality can be calculated using formulas that can be easily applied to raw review data in large datasets. For example, measures of our three tested constructs (comment amount, mean comment helpfulness, and rating accuracy) can be calculated using formulas in spreadsheets based upon available review data rather than laborious hand-coding approaches or complex Natural Language Processing (NLP) techniques that typical require training using subsets of hand-coded data. Next, we present a brief review of prior work on each of these three specific review quality constructs.

1.1.1. Comment amount

Although peer review systems often apply some requirements to the amount of feedback produced (e.g., the required number of reviews per assignment, minimum word counts for each comment and a minimum number of comments per review rubric), many systems do not have such features and, even in the systems that do have requirements, students are typically able to go beyond the minimums (e.g., submit bonus reviews, provide very long comments, provide more comments within a rubric dimension). Some prior work on the amount of feedback produced has focused on *frequency*, either in terms of the number of reviews completed or the numbers of comments provided within a review or across reviews (Zong et al., 2021b; Zou et al., 2018). Other researchers have examined the feedback amount regarding comment length, particularly the number of words in the comments produced (e.g., Jin et al., 2022; Patchan et al., 2018). Sometimes this more granular measure is integrated with frequency, such as in the total number of words produced across comments (Zong et al., 2021b). Since the relationship of comment length to comment quality is unlikely to be linear (e.g., very long comments could be confusing; very brief comments could be worthless), some researchers have counted the number of low-quality comments, defined as comments having fewer than ten words, and the number of high-quality comments, defined as comments having more than 50 words (Patchan et al., 2018; Zong et al., 2021a, 2022a). These specific thresholds were validated on the basis of correlations with student ratings of comment helpfulness (Zong et al., 2021a).

In this paper, we focus on the number of long comments (i.e., the number of comments with more than 50 words) because it integrates both the frequency and length dimensions and addresses concerns about non-linear relationships to comment quality. Further, it is predictive of student learning; that is, producing more long comments defined in this way is a robust predictor of task improvements by the reviewer in future assignments (Zong et al., 2021a; Jin et al., 2022). It has also been found to be sensitive to differences in the task-reviewing environment (Patchan et al., 2018). In addition, it has previously been found to increase and decrease as a function of prior experiences with peer feedback (Zong et al., 2022a). Finally, Zong et al. (2022b) found this measure had good reliability (Cronbach's alpha 0.89), with higher reliability than a measure based upon the percentage of long comments produced.

1.1.2. Comment accuracy

Researchers have approached comment accuracy from a wide variety of perspectives: expert coding of comment potential impact (Wu & Schunn, 2021a), expert coding of the number of errors detected (Van Steendam et al., 2010), expert coding of number of meaningful comments (Cui et al., 2021; Kobayashi, 2020), expert coding of student comments on the helpfulness of their received comments (Nelson & Schunn, 2009; Wu & Schunn, 2020), and student ratings about the helpfulness of the comments (Patchan et al., 2018; Zong et al., 2021a). The last approach is especially viable at scale when applied to data collected using the various systems that require students to provide such ratings are part of the review process, as in the current study context. A number of studies have looked at students' helpfulness ratings to find what comment features are associated with helpfulness ratings (e.g., Yallop & Leijen, 2018). One study focused on helpfulness at the students' level and showed that changes (both increases and decreases) in comment helpfulness from one assignment to the next could be predicted by the number and length of feedback provided in the prior assignment (Zong et al., 2021b).

1.1.3. Rating accuracy

Rating accuracy has been examined in terms of measures based upon intra-rater reliability, expert-student comparisons, and inter-rater reliability (Cho et al., 2006; Jonsson & Svingby, 2007; Suen, 2014; Vanderhoven et al., 2015). Intra-rater reliability represents the consistency of using a rating system within one peer reviewer from assignment to assignment (Kayapinar, 2014; Sherrard et al., 1994; Xiong & Schunn, 2021); for example, do students become more or less lenient over time? However, intra-rater reliability might not be a major concern when reviewers are supported by a rubric (Jonsson & Svingby, 2007). Expert-student comparisons are often used to assess the validity of student ratings (Falchikov & Goldfinch 2000; Rushton et al., 1993; Stefani, 1994), and several *meta*-analyses have found moderately-sized average correlation coefficients between peer ratings and expert scores (0.69 for Falchikov & Goldfinch, 2000; 0.63 for the *meta*-analysis by Li et al., 2016; 0.47 for Panadero et al., 2013). However, expert evaluations are often not available, particularly in larger datasets.

Inter-rater reliability represents the agreement among different reviewers on the same document (Cho et al., 2006; Suen, 2014). Since inter-rater reliability is a strong correlate of variations in the validity of student scores (Xiong & Schunn, 2021) and it is always possible to calculate, we focus on inter-rater reliability. Inter-rater reliability values, defined as rating consistency, typically range from 0.3 to 0.6 (Cho et al., 2006; Paré & Joordens, 2008; Zhang et al., 2020). This variation in inter-reliability has been found to be influenced by students' prior experiences. For example, adding a comment accountability feature (asking students to report the helpfulness they received from their peers) increased students' inter-rater reliability (Paré & Joordens, 2008; Patchan et al., 2018). The widely used CPR (Calibrated Peer Review) system was specifically developed to provide students with initial practice with feedback on a rubric to improve students' inter-rater reliability, defined as absolute deviations, and there have been some reports that it was successful in this regard (Carlson & Berry, 2008). We focus on agreement measures (i.e., deviations among raters in rating each document) because consistency measures (i.e., correlations across raters in the relative ratings of documents) require more ratings than an individual will typically produce.

1.1.4. Selected review quality measures

In sum, we will examine review quality at the reviewer level in terms of three constructs: comment amount (i.e., number of provided long comments), comment accuracy (i.e., helpfulness of comments), and rating accuracy (i.e., rating agreement). And the first part of our study investigates the stability and separability of the three constructs as the reviewer level, alongside the reliability of the measures. With these properties established, it becomes possible to study how reviewers self-regulate these characteristics, the second part of our study.

1.2. Self-regulation of feedback quality

More broadly, self-regulated learning (SRL) has seen significant advancements since the 1970s, conceptualizing self-regulated learners as active participants in their own cognitive and metacognitive processes, which includes planning, goal setting, and the use of task-related strategies and self-explanation to achieve their learning objectives (Schunk

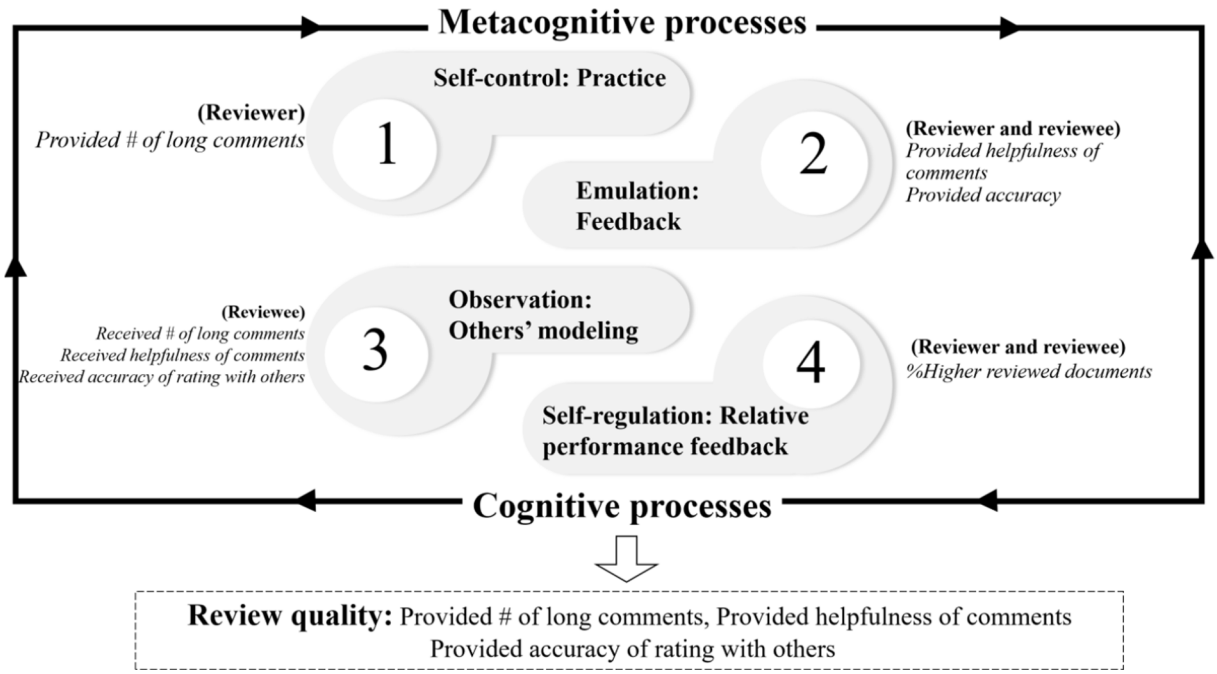


Fig. 1. Seven specific experiences with peer feedback (in italics) involving both reviewer and reviewee experiences (noted in parentheses) that can be organized into four conceptual sources (bold black font). Cognitive and metacognitive processes are applied to this information in a cylical process, resulting in the self-regulation of review quality.

& Greene, 2017; Zimmerman, 1989). Numerous theoretical perspectives on self-regulated learning have emerged with perspectives ranging from cognitive-developmental (Thoresen & Mahoney, 1974), to meta-cognitive and cognitive (Zimmerman & Schunk, 2011), and to social and motivational (Schunk, 2012). Despite their differences, these perspectives share common features, such as viewing SRL as a dynamic and cyclical process that comprises feedback loops, with self-regulated learners setting goals and metacognitively monitoring their progress toward them (Lord et al., 2010). Further, self-regulated learning is thought to respond to information obtained through monitoring and external feedback in ways they believe will help them achieve their goals, such as by working harder or altering their strategies. Prior performance and experience have been found to be critical for helping students shift towards more productive SRL behaviors in the next learning cycles (Raković et al., 2022; Zhang et al., 2022).

From a lens of self-regulated learning, during peer review, students go through the steps of observing peers' work, providing feedback, reading peers' comments, and reflecting on their own document, each of which provides opportunities for feedback on and practice with peer review (Zong et al., 2021b). As noted above, such feedback-practice loops are a critical part of SRL. To further the mapping between SRL and peer feedback, some SRL theories have framed students as adapting through multiple levels of experiences or training (e.g., Zimmerman, 2013), all of which are present in peer feedback: by **self-control** through practice and by **emulation** based upon social feedback as a reviewer, by **observation** from others' modeling as a reviewer and reviewee, and by **self-regulation** through relative performance feedback both as a reviewer and reviewee. We argue that these experiences during peer review contain aspects that map onto each of these four conceptual sources of self-regulation and, therefore could shape changes in peer review quality as a kind of self-regulation process through cognitive and metacognitive processes applied these sometimes complex sources of information (see Fig. 1). Note that this self-regulation process can produce both increases and decreases in review quality. For example, information derived from observation or social feedback can be interpreted by students to suggest that lower feedback quality is expected in this context.

1.2.1. *Practice*

Practice is perhaps the most common self-regulation method, and practice that focuses on higher-quality behaviors with higher cognitive engagement is likely to produce more improvement (Fredricks et al., 2004). In the context of peer review, Baars et al. (2020) and Zhu and Carless (2018) argued that providing long comments benefits reviewers because longer comments imply the reviewer is more cognitively engaged. Providing more long comments has been associated with improvements in task performance (Zong et al. 2021a) and changes in the helpfulness of reviews (Zong et al., 2021b). Thus, it is reasonable to infer that providing more long comments in one assignment should be associated with increases in various aspects of feedback quality in the next assignment.

1.2.2. *Feedback*

Feedback is widely viewed as useful for learning and self-regulation (Van der Kleij et al., 2015), and a recent *meta*-analysis found a moderate positive effect of self-regulation feedback on metacognitive and resource management strategies as well as on motivation (Theobald, 2021; $0.26 < g < 0.34$). Feedback on review quality is, therefore, likely to influence review quality. A number of peer reviewing techniques include some back evaluation component, defined as "the feedback that an assessee provides to an assessor about the quality of the review" (Luxton-Reilly, 2009, p. 226). It could take the form of recognizing particularly helpful comments and rating accuracy (Patchan et al., 2018; Zong et al., 2022a). That is, providing reviewers with information about how helpful their peers found their comments and providing information about their rating accuracy are the two common forms of feedback on review quality. Such feedback on review quality is likely to influence feedback quality because they lead students to reflect on their review quality (Misiejuk & Wasson, 2021). Producing reviews that receive higher helpfulness ratings from comment recipients has also been associated with growth in task performance (Zong et al., 2021b). But it is unclear whether recognition for providing helpful comments would lead to providing more long comments and accurate ratings indirectly through improvements in underlying task performance. Several systems provide information on rating accuracy to students (e.g., Peerceptiv,

Table 1
For each year of the course, # of participating students self-reported mean age, % female, and % reporting each race/ethnicity among those who reported demographics.

Course Year	N	Mean age	% Female	Race/Ethnicity			
				% Asian	% Black	% Latinx	% White
2015	696	21.0	45%	88%	0%	7%	5%
2016	689	20.5	43%	68%	1%	15%	16%
2018	767	–	–	–	–	–	–

Note. - = not recorded.

MobiusSLIP, CPR, Kritik), presumably because providing such feedback information improves rating accuracy. However, this assumption, though plausible, has not been directly assessed. Further, it is unclear whether providing reviewers with information on their rating accuracy will also lead to growth in the amount of commenting and higher comment accuracy; but if it does occur, it theoretically should be a positive relationship (i.e., positive feedback should produce increases in comment amount and comment accuracy).

1.2.3. Others' models

Often called modeling in the research literature, simply observing outputs from others can shape students' behavior by providing descriptive information (e.g., learning about new strategies for completing a task; Bandura et al., 1966) and normative information (i.e., learning what behaviors are expected; Press & Dyson, 2012; Stewart & Plotkin, 2013). In peer review, the peer feedback students receive in prior assignments can serve as a model for how reviews should be provided in future assignments. Receiving more comments has been associated with later providing more comments, and receiving fewer comments has been associated with later providing fewer comments (Zong et al., 2022a). Similar patterns were observed for the helpfulness of received comments and the helpfulness of later provided comments (Zong et al., 2022b). At the same time, there is the potential for negative motivational effects: receiving more comments, especially if they tend to be critical, may demotivate students and they may then provide lower quality feedback in the next assignment. Further, it is unclear whether receiving more (or less) accurate ratings would lead to providing more (or less) accurate ratings.

1.2.4. Relative performance

Information about one's own overall performance level in a domain, especially in comparison to others' performance, can have substantial effects on increasing or decreasing motivation (Margolis & McCabe, 2006). In peer review, students receive this information through the contrast in quality they observe between their own documents and those of their assessees. Many studies have examined whether students are better off in terms of task learning from receiving or providing feedback to students who are at similar, lower, or higher performance levels (Alqassab et al., 2018; Dmoshinskaia et al., 2021b, 2022; Tsivitanidou et al., 2018). No prior studies have examined whether reviewing quality, in particular, is influenced by this relative contrast. However, given the broad motivational impacts of relative performance information, particularly for self-efficacy (Pajares, 2003), and the role of self-efficacy in being willing to provide substantial feedback, it is reasonable to hypothesize that relative performance information (own document quality vs. assessed document quality) will impact future review quality.

1.3. The present study

Based on the literature on peer review quality, we investigated two major questions regarding: 1) the existence of separate constructs of review quality at the level of reviewers within the assignment; and 2) what self-regulation information sources (i.e., experiences with peer feedback) are positively or negatively associated with subsequent changes over time in each construct of review quality. The experience with peer feedback specifically means what students experienced in their immediately prior assignment. The study uses a large peer review dataset involving naturally collected data from the same large enrollment biology course, which includes four peer review assignments each semester across three semesters. In particular, the study tests eight specific hypotheses associated with the two main research questions:

RQ1: Are the three examined feedback quality constructs meaningfully conceptualized at the level of reviewers within an assignment?

H1: Comment amount, comment accuracy, and rating accuracy can be reliably measured at the reviewer level within an assignment.

H2: Comment amount, comment accuracy, and rating accuracy are each semi-stable dimensions of review quality (i.e., meaningful reviewer characteristics).

H3: Comment amount, comment accuracy, and rating accuracy are separable constructs of review quality.

RQ2: Are the four self-regulation information sources positively or negatively associated with changes in the three constructs of feedback quality?

H4: *Practice* comment amount (number of long comments provided in the prior assignment) should be positively associated with changes in review quality (both comments and rating accuracy).

H5a: *Feedback* on comment accuracy (helpfulness ratings provided in the prior assignment) should be positively associated with changes in review quality (number of long comments and rating accuracy).

H5b: *Feedback* on rating accuracy (level of accuracy provided in the prior assignment) should be positively associated with changes in feedback quality (number of long comments and comments accuracy).

H6: Other students' *models* of review quality (received long comments, received helpfulness ratings, received rating accuracy in the prior assignment) may be positively or negatively associated with changes in review quality (number of long comments, comments accuracy, and rating accuracy).

H7: *Relative performance* (the percentage of reviewed documents in the prior assignment that were stronger documents than the document the reviewer produced at the previous assignment) may be positively or negatively associated with changes in review quality (number of long comments, comments accuracy, rating accuracy).

Note that the specific hypotheses in RQ2 explore the role of available experience measures within each of the four types of self-regulation information on the three focal constructs of feedback quality but exclude the cases in which the prior experience measure is also the baseline feedback quality measure. For example, practice (prior comment amount) can only be explored as influencing comments and rating accuracy, but not for testing the relationship with the comment amount itself.

2. Methods

2.1. Participants and context

Participants were 2,092 undergraduate students attending a large public research university on the west coast of the US. The students were enrolled in one of three offerings of the same upper-level Biology course for biology majors taught by the same instructor in 2015, 2016, and 2018. This upper-level required course for biology majors aimed to develop students' scientific writing in biology, which included developing a better understanding of scientific articles and general communication of biological research to other audiences. Across the different offerings, variations of assignments were implemented, sometimes analyzed different kinds of published papers were, and sometimes students practiced different sections of a research paper (e.g., methods, results, discussion). These offerings of the course were selected because they all used the same general peer feedback tool (Peerceptiv) and general approach (i.e., individual rather than group-submitted

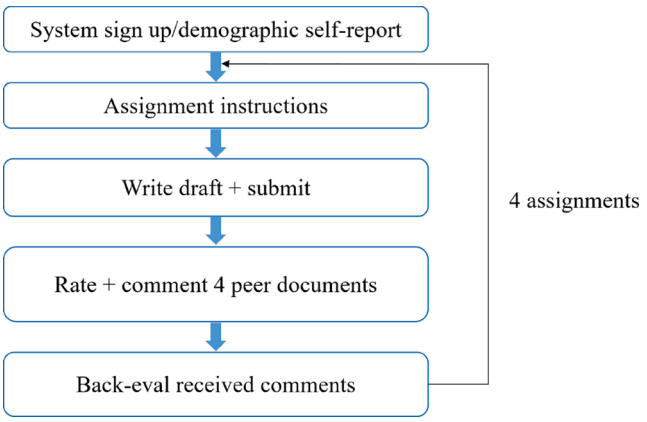


Fig. 2. The peer assessment process within each course offering.

Table 2
The specific measures for each construct and their definitions.

Construct	Measure	Definition
Assignment number	J	The sequence of assignments within a course (from 1 to 4)
%Higher reviewed documents	$\%Higher_J$	Percentage of reviewed documents that were of higher quality than own document.
Characteristics of feedback and ratings provided by a given student on the J^{th} assignment		
# of long comments	$Provided\ \#Long_J$	The number of long provided comments (#words > 50).
Helpfulness of comments	$Provided\ Helpfulness_J$	The mean helpfulness rating for provided comments.
Accuracy of ratings with others	$Provided\ Accuracy_J$	The mean accuracy with the average rating given by other reviewers for the provided ratings.
Characteristics of feedback and ratings received by a given student on the J^{th} assignment		
# of long comments	$Received\ \#Long_J$	The number of received comments that were long (#words > 50).
Helpfulness of comments	$Received\ Helpfulness_J$	The mean helpfulness rating for received comments.
Accuracy of ratings with others	$Received\ Accuracy_J$	The mean accuracy across reviewers for the received ratings.

assignments) and included enough data (students and assignments) in each offering to conduct multiple regression within an assignment. Yet, the replication and generality of findings across different assignment structures could be tested. The data has not been analyzed for peer review quality before.

Table 1 shows the self-reported demographic details of students across the three offerings of the Biology course; 40% of students reported demographic details for 2015 and 2016. After 2017, accounts were automatically created in the peer review system through a direct connection with the learning management system, so students no longer provided demographic information at account setup. Based upon university demographic trends, it is likely that the demographics in 2018 closely mirrored demographics from the prior years. The high proportion of students reporting Asian race/ethnicity reflects a combination of the higher proportions of Asian students in that region in the US attending selective universities and choosing to major in a health-related STEM discipline.

2.2. Materials

All students were required to submit assignments and complete peer reviews through Peerceptiv, which has been widely used in university contexts across disciplines around the world for online peer review (https://peerceptiv.com; Schunn, 2016). First, students submitted their documents for peer review under a pseudonym. Then, students anonymously assessed four randomly-assigned peers' documents using instructor-provided rubrics (see Fig. A1) that directed students to focus on specific aspects of the assignment (e.g., strengths, weaknesses, and organization) while viewing the submitted document. Students also had to submit at least one comment per reviewing dimension, but no minimum or maximum comment length was required. Interleaved with comments, students also needed to provide peer assessment on 1-to-7 scales specific to each reviewing dimension (and potentially multiple ratings per dimension).

At the end of the reviewing period (typically one week), authors needed to rate the helpfulness of each of their received comments on a 1-to-5 scale (called back-evaluations) and include an optional explanation for their rating. Reviewers could then see the back-evaluations of their provided comments received alongside back-evaluations given to comments provided by other reviewers on that same document (see Fig. A2). These back-evaluation ratings served as the basis for one measure of comment quality but also acted as an incentive system to encourage reviewers to provide critical and constructive comments in their reviews. In addition, after the rating phase, reviewers could view their rating accuracy (via a supporting graph that shows the ratings they provided against the means others provided). Thus, helpfulness and rating accuracy are characteristics of the reviewer for which the reviewer receives feedback within the system on each assignment. The procedure for one class is shown in Fig. 2.

2.3. Measures

All measures were derived from data automatically collected within the online system (see Table 2 for an overview of all the measures, associated variable names, and their definitions). The Human Research Protection Office approved research conducted on this anonymized data at University of Pittsburgh. The data were organized to enable separate multiple regression analyses for each assignment in each course offering, and then *meta*-regression was used to calculate mean effect sizes; this analytic approach is more robust than the previously-used approach of applying multiple regression to the entire course or multiple courses with a linear predictor variable for assignment number and dummy codes for courses (e.g., Zong et al., 2021a), since *meta*-regression applied at the assignment level allows for the possibility of different regression patterns in different assignments. Further *meta*-analysis provides formal measures of variability in effects across assignments and course offerings in the form of heterogeneity estimates, separating variability due to uncertainty in estimates from meaningful variation across contexts.

The outcome variables were three review quality measures (two involving comment quality and one involving rating quality) on the J^{th} assignment. The combination of the likely influence of the reviewed document and the small number of reviewed documents by each reviewer will likely limit the observed reliability that is found; that is, relatively modest reliability values are likely to occur. It is important to note that standards for what are acceptable levels of measure reliability depend both upon the purposes of measurement and the challenges of the context. For example, measures that are intended to support important decisions about an individual, such as career or educational placements, should have very high reliability. By contrast, measures only used for research purposes can have lower levels of reliability, and their acceptability can depend upon the amount of data available. For example, large national survey studies often have measures based upon single items to enable sampling more constructs while still having high response rates (i.e., avoiding painfully long surveys). The increase in measurement noise is acceptable because of the statistical power afforded by the sample size. Similarly, the challenges of the context also influence standards. For example, a number of neurophysiological measures, like those using fMRI and ERP are very indirect and require aggregations of many thousands of data points to overcome the inherent noise of the measurement and yet are widely accepted forms of research

(Seghier et al., 2019). By contrast, survey research expects a reliability of 0.8 for survey instruments (Hair et al., 2010), and knowledge assessment research expects reliability of 0.7 (Peters et al., 2012). In peer review research, measure reliabilities are rarely above 0.7 (Jonsson & Svingby, 2007; Van Zundert et al., 2010; Popken, 2020), and sometimes reliabilities are considered poor only if they are below 0.4 (Zhang et al., 2020).

The same measure on the prior assignment was used as a baseline within a time-series analysis for the regression models examining changes in these review quality outcome variables. Key potential predictors of growth involved various experiences in the prior (i.e., J-1) assignment (e.g., how much feedback was provided or received). Only the prior assignment was considered because including greater lags significantly reduces the total amount of data when there are only four assignments. In addition, prior work established a close relationship with experiences on the J-1 assignment but not with the J-2 or J-3 assignment (Zong et al., 2021a). Since the 1st assignment in a course had no prior experience data, this approach produced nine datasets (3 assignments × 3 courses).

%Higher reviewed documents. The variable reflects the percentage of reviewed documents that were of higher quality than the reviewer's own document for that assignment. The randomly assigned documents to review are often of varied quality, but by-chance or because the reviewer is relatively low or relatively high in performance, a given reviewer for a given assignment might have received documents that were primarily of higher quality than their own document or primarily of lower quality than their own document. The overall quality of documents (own and others) was calculated using the mean of all ratings received for each document. The mean rubric-based ratings across multiple peers tend to have high reliability and good validity for peer feedback in general (Li et al., 2016) and for Peerceptiv in particular (Patchan et al., 2018; Wu & Schunn, 2021a). Further, the effect of this variable is conceptualized in motivational terms. Thus the relative quality difference as perceived by students is more relevant than the relative quality difference as perceived by the instructor or experts.

 # Long provided and received. The Peerceptiv interface has no required minimum or maximum length for the open-ended peer feedback comments. Thus, students can provide comments consisting of a single word or several paragraphs in the textbox. Prior research on peer feedback (e.g., Nelson & Schunn, 2009; Wu & Schunn, 2020) has found that compared to short comments, long comments can provide more useful information to the authors, such as pointing out the problems, explaining why it is important, giving an example, evaluating the quality, and giving the useful suggestions. Based upon previous research, we defined long comments as those having at least 50 words. Compared to other possible threshold values, this particular threshold is the better predictor of what authors consider to be helpful comments (Zong et al., 2022b) and it also results in a good predictor of student learning from provided comments (Zong et al., 2021b). We calculated the number of long comments across all reviews received by a student for assignment *J* (*Received #Long*$_J$) or all the reviews provided by a student for assignment *J* (*Provided #Long*$_J$).

Helpfulness of comments provided and received. As described in the materials section, students were required to judge the helpfulness of the comments they received on a 1-to-5 Likert scale, and these judgments are typically based upon understandability and agreement (Nelson & Schunn, 2009). In the two examples shown below, backevaluation comment 1 justified giving a helpfulness rating of 5, whereas backevaluation comment 2 justified giving a helpfulness rating of 1.

Backevaluation comment 1: You're right, I hadn't considered if adding that statement would result in an unscientific tone. I just wrote it in that manner to demonstrate the problem at hand without really thinking of professionalism.

Backevaluation comment 2: While I appreciate the positive reinforcement, this critique does not tell me how to improve and is very vague. No detail is really present.

We calculated the mean helpfulness across all comment dimensions and reviews received by a student for assignment *J* (*Received Helpfulness*$_J$) or all the reviews provided by a student for assignment *J* (*Provided Helpfulness*$_J$).

Accuracy of ratings provided and received. In the peer assessment rubrics, students were required to rate the document's quality using 1-to-7 scales for each dimension. We first calculated an overall score for a document by a reviewer (i.e., the mean of the ratings across dimensions). And each document was given an overall score based upon the mean score across reviewers. Then we normalized all scores based upon raw ratings (overall document score and reviewer assessment of a document) to have a mean of 85 and standard of 10, which corresponded to the grade curving scheme in the course. After that, we used the absolute value of the difference between the document grade and the reviewer's normalized rating score as the disagreement score. After observing the range of observed disagreement scores, accuracy scores were computed as 15 (the highest score) minus the disagreement score (i.e., theoretical max of 15 and min 0, with 15 representing no disagreement). We finally calculated mean accuracy across all reviews received by a student for assignment J (*Received Accuracy*$_J$) or all the reviews provided by a student for assignment *J* (*Provided Accuracy*$_J$). Pilot work explored an alternative method to calculating rating accuracy: the correlation between the student's provided ratings and the mean ratings provided by other students on those same documents and rating rubrics. This approach is directly implemented in Peerceptiv, and it is similar to consistency, rather than agreement, type of interclass correlation (McGraw & Wong, 1996). However, pilot regression analyses using the consistency measure showed lower temporal stability and inconsistent/weak predictors of growth. Therefore, the agreement rather than the consistency approach to measuring rating accuracy is used in the reported analyses.

2.4. Analysis

All analyses were conducted in R. To test the three hypotheses associated with RQ1, intra-class, and inter-class correlations were calculated within each assignment and then *meta*-correlation analyses were conducted using the "meta" package in R to determine the average correlation magnitude and its statistical significance. To test the five hypotheses associated with RQ2, multiple regressions were implemented for each assignment, and then *meta*-regression was applied using the "metafor" package in R to calculate average regression coefficients and their statistical significance. Meta-regression, rather than multilevel modeling, is also suitable for multilevel data, and it is especially useful when the same data collection methodology is used across distinct sites, but a different pattern of results is seen in different sites (Benton, 2014; Pastor & Lazowski, 2018). Significant heterogeneity was expected because *meta*-analyses of the effects of peer review on various outcomes always find large heterogeneity of effects (Falchikov & Goldfinch, 2000; Huisman et al., 2019; Li et al., 2016; Li et al., 2020; Panadero et al., 2013).

For each correlation coefficient or predictor beta weight, the *meta*-correlations / *meta*-regressions provided a mean, *p*-value, and 95% confidence interval. Before conducting each correlation and multiple regression, outliers in each continuous predictor were replaced with the closest non-outlier value because this approach conservatively mitigates the influence of abnormally large (or small) measurements on the normalized distance (Costa, 2014; Van Selst & Jolicoeur, 1994). In addition, data were treated as missing and deleted on purpose (for at most 8% of data): 1) when students failed to submit a document in the prior assignment, the received review quality measures were treated as missing; and 2) when students failed to complete any reviews, the provided review quality measures were treated as missing. In the time-lagged regressions, a student was excluded if they were missing either a predictor or an outcome (the received/provided review quality measures were missing). This listwise deletion approach left most of the data

Table 3
Meta-correlation estimates across courses and assignments of mean correlations among predictors (upper), outcomes (lower right), and between predictor and outcomes (bottom left). Grey cells represent the lag-1 stability of each variable. Bold values are those whose absolute value is greater than 0.15.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1 %Higher$_{J-1}$	**0.28*****									
2 Provided #Long$_{J-1}$	**−0.25*****	**0.68*****								
3 Provided Helpfulness$_{J-1}$	**−0.19*****	**0.53*****	**0.42*****							
4 Provided Accuracy$_{J-1}$	0.08***	−0.02	0.10***	**0.22*****						
5 Received #Long$_{J-1}$	0.09***	−0.01	−0.03	0.01	0.03					
6 Received Helpfulness$_{J-1}$	−0.07***	−0.07***	0.04*	0.08***	0.11***	**0.49*****				
7 Received Accuracy$_{J-1}$	**−0.41*****	0.10***	0.09***	0.02	−0.05*	0.09***	0.09***			
8 Provided #Long$_{J}$	**−0.21*****	**0.68*****	**0.42*****	−0.005	−0.001	−0.06***	0.09***	**0.68*****		
9 Provided Helpfulness$_{J}$	**−0.18****	**0.43*****	**0.42*****	0.06**	0.003	0.04**	0.10***	**0.52*****	**0.42*****	
10 Provided Accuracy$_{J}$	−0.01	0.01	0.07**	**0.22*****	−0.0003	0.06***	0.04*	−0.002	0.11***	**0.22*****

Notes. *=p <.05, *=p <.01, ***=p <.001.

intact. Moreover, since students who missed one task or review tended to miss multiple measures, imputation methods were deemed unlikely to provide substantial improvements to the analysis.

For the purposes of examining the reliability of each outcome measure (H1), scores on each measure were re-calculated at the level of an individual review (rather than at the level of all reviews provided), and we calculated intra-class correlation coefficient as the measure of reliability based upon the first four reviews, the minimum number of reviews that each student provided. Formally within the McGraw and Wong (1996) framework, this was the ICC(C, k), consistency-type reliability of the aggregate measure of *k* ratings, similar to a Cronbach's alpha, but using a one-way random model because different reviewers rated different documents. The reliability of each measure was calculated for each assignment in each course, and then *meta*-correlation was used to calculate the mean ICC and 95% CI for each measure. Given that the goal was to test whether measures were sufficiently reliable for quantitative analysis in large datasets, minimum acceptable reliability was set at 0.4, and strong reliability was set at 0.7.

To address H2 and H3 in Research Question 1, we examine the time-lagged correlations (H2) within the three review quality measures and the correlations among the three measures of review quality (H3). Again, *meta*-correlation was used to calculate each mean correlation and 95% CI across courses and assignments. A feedback quality measure was considered semi-stable (H2) if the mean time-lagged correlation of that measure was within the range of 0.2 to 0.8. A feedback quality measure was considered separable (H3) if no mean correlation with another feedback quality measure was above 0.8.

To test H4–H7 within Research Question 2, we use multiple regressions involving time-lagged models, with the three review quality measures (*Provided #Long*, *Provided Helpfulness*, *Provided Accuracy)* on assignment *J* being predicted by *%Higher*, *Provided #Long*, *Provided Helpfulness*, *Provided Accuracy*, *Received #Long*, *Received Helpfulness*, and *Received Accuracy* for assignment *J-1*. The outcome variable *Provided #Long$_J$* is a count variable, which tends to have a strong positive skew. In such situations, Poisson regression or Negative Binomial regression is preferred over linear regression (Grogger & Carson, 1991). In addition, count variables can occasionally have an excess number of zeros, which can be addressed using zero-inflation models, which split the regression into two components: predicting zero and then predicting the number for non-zero cases. The AICs (Akaike Information Criterion) showed better fits for Zero-Inflated Negative Binomial regression, and therefore this approach is reported here, implemented using the *zerofinl* command in the "pscl" package in R. The other two dependent variables (*Provided Helpfulness*, *Provided Accuracy*) were continuous variable with roughly normal distributions in each assignment, and therefore we used linear regression models. Meta-regression was used to calculate estimated means for each predictor's beta weight. Note that predictors and outcome measures were standardized in the analyses to allow for a direct comparison of effect sizes. Violin box plot for each variable across assignments in each of the three courses are presented in Appendix Table A2.

3. Results

3.1. Are the three examined feedback quality constructs meaningfully conceptualized at the level of reviewers within an assignment?

3.1.1. H1: Reliability

Focusing on the reliability of each feedback quality measure, the *meta*-correlation analyses revealed that all three feedback quality measures met the minimum reliability threshold (0.4), and one measure showed strong reliability: *#Long* 0.89, 95%CI [0.84; 0.92], *Helpfulness,* 0.53, 95%CI [0.49; 0.57], and *Accuracy,* 0.48, 95%CI [0.44; 0.51]). Appendix Table A1 shows the intra-class correlation coefficients for each feedback quality measure in every assignment; there was no general pattern across courses or assignments when reliability tended to be higher. It is important to note that in the context of peer review, measures reliabilities are rarely above 0.7 and sometimes above 0.4 is taken as acceptable in studies involving large datasets (e.g., Jonsson & Svingby, 2007; Zhang et al., 2020). The primary cause of lower reliability of reviewing quality measures as reliability is typically measured (i.e., consistency across reviews) is that reviewing behaviors are jointly influenced by reviewer characteristics and document characteristics. Thus, with the contextual expectations regarding measure reliability, Hypothesis 1 is strongly supported for comment length and moderately supported for comment helpfulness and rating accuracy.

3.1.2. H2: Semi-stable

The gray cells in Table 3 show the mean estimated lag-1 correlations for each predictor and outcome variable. The temporal stability of the feedback quality measures directly relevant to H2, is shown in the bottom right. The two measures of comment quality are moderately stable from one assignment to the next (*#Long* 95%CI [0.65, 0.71] and *Helpfulness* 95%CI [0.37, 0.47]), whereas the measure of rating quality (accuracy 95%CI [0.18, 0.27]) has more modest temporal stability. All three measures fell within the required range to be considered semi-stable. It is possible that the much higher stability of the comment length variable was partially due to the higher reliability of that measure. However, it is interesting that rating accuracy was substantially less stable than comment helpfulness even though they had similar measure reliability.

3.1.3. H3: Separable

The off-diagonal values in the lower right area of Table 3 shows that the number of long comments and comment helpfulness were moderately correlated with one another (95%CI [0.46, 0.58]) but that the correlations of comment quality with rating quality were very small: #Long 95%CI [-0.05, 0.05] and Helpfulness 95%CI [0.08, 0.15]. All of these correlation values were sufficiently small to support Hypothesis 3; the three review quality dimensions were separable.

Table 4
Meta-analysis results for the overall effect across assignments of each core predictor of #Provided Long, *Provided Helpfulness* and *Provided accuracy* along with heterogeneity of effects across assignments (statistically significant overall effects are in bold).

	#Provided Long					Provided Helpfulness					Provided Accuracy				
	Standardized β			Heterogeneity of Effect		Standardized β			Heterogeneity of Effect		Standardized β			Heterogeneity of Effect	
	Mean	*95% CI*		p	I^2	p	Mean	*95% CI*		p	I^2	p	Mean	*95% CI*		p	I^2	p
Baseline	**0.32**	0.24	0.40	<0.001	98%	< 0.001	**0.25**	0.19	0.30	<0.001	64%	< 0.001	**0.21**	0.16	0.27	<0.001	66%	< 0.001
#Provided Long$_{J-1}$							**0.29**	0.23	0.35	<0.001	100%	< 0.001	−0.01	−0.04	0.02	0.40	76%	< 0.001
Provided Helpfulness$_{J-1}$	**0.08**	0.04	0.11	0.001	87%	< 0.001							**0.05**	0.01	0.09	0.03	0%	1
Provided Accuracy$_{J-1}$	0.00	−0.01	0.01	0.86	12%	0.23	**0.04**	0.00	0.08	0.049	98%	< 0.001						
#Received Long$_{J-1}$	0.01	−0.02	0.04	0.35	84%	<0.001	0.02	−0.01	0.04	0.22	98%	< 0.001	−0.01	−0.03	0.01	0.37	9%	0.42
Received Helpfulness$_{J-1}$	**−0.02**	−0.04	−0.003	0.03	64%	0.003	**0.05**	0.02	0.07	<0.001	1%	0.376	**0.04**	0.03	0.06	<0.001	0%	1
Received Accuracy$_{J-1}$	−0.01	−0.02	0.01	0.44	45%	0.04	0.02	−0.01	0.05	0.17	95%	< 0.001	0.03	−0.02	0.07	0.26	37%	0.07
%Higher$_{J-1}$	**−0.03**	−0.05	−0.004	0.03	79%	< 0.001	**−0.05**	−0.08	−0.03	0.001	0%	0.942	−0.01	−0.07	0.05	0.73	0%	1

3.2. Are the four self-regulation information sources positively or negatively associated with changes of the three constructs of feedback quality?

Table 3 also contains important information related to Research Question 2 (Hypotheses 4–7). In the lower left of the table are the mean correlations of each of the predictors with each of the three outcome variables. Future comment quality is significantly associated with all of the aspects of prior provided comment quality at moderate-to-substantial levels and with received comment quality at much smaller levels. Future rating quality is mildly associated with one aspect of prior provided comment quality and two aspects of received review quality.

Since each review quality outcome had significant correlations with multiple predictors and all of the predictors were significantly correlated with multiple other predictors (see the upper part of Table 3), multiple regression was needed to tease apart the unique relationship of each predictor with future review quality. Conveniently, most of the mean correlations among the predictors were small to medium (none greater than 0.53), and therefore multicollinearity problems were not expected. Indeed, an examination of variance inflation factors (VIFs) confirmed that there were no multicollinearity problems (VIFs < 2 for every assignment-specific regression). In addition, to examine predictors of change in each dimension of review quality, the multiple regressions needed to control for prior review quality in the time-series analysis. Since the lag-1 correlations (gray cells) for the review quality variables were between 0.2 and 0.7, at least half of the variance in the variables could be conceptualized as changes in reviewing quality from one assignment to the next that could be explained by prior experiences.

Table 4 presents the full *meta*-regression results for every predictor, including mean and 95% CI of the estimated standardized effect and the size and statistical significance of between course/assignment heterogeneity of the estimated effect. Fig. 3 summarizes the key *meta*-regression findings for RQ2: the extent to which each key predictor significantly explained the changes in each of the three review quality measures. The predictors are organized by the theoretical construct they most closely align. The figure also shows that the baseline (lag-1) variables accounted for a roughly similar amount of variance for all three review quality variables, and they were positively significant. This level of baseline effect is consistent with a semi-stable characteristic of the reviewers that is also influenced by recent experiences. Fig. 4 shows the approximate effect size by plotting the marginal means of each measure of reviewing quality as a result of having high or low levels of a given prior experience (holding all other experiences constant).

3.2.1. H4. The amount of practice is positively associated with changes in feedback quality.

As shown towards the left side of Fig. 3 and the second row of Table 4, *Provided #Long$_{J-1}$* was a significant predictor of change in provided helpfulness, with a large mean estimated effect of 0.29 and a 95%CI of [0.23, 0.35], providing strong confidence that the mean effect is substantial. By contrast, the mean estimated effect for predicting change in provided accuracy is only -0.01 with a 95%CI of [-0.04, 0.02], providing strong confidence that the mean effect is, at best, very small. Thus, the practice effect (H4) is strongly supported for one dimension of review quality (helpfulness of comments) and strongly ruled out, at least in general, for another dimension of review quality (accuracy of ratings). At the same time, Table 4 shows that there is also very large and statistically strong heterogeneity of both practice relationships across courses and assignments, suggesting contextual details are important in shaping the amount of learning that is observed from practice experiences. For example, practice effects might be occurring for the accuracy of ratings in some assignments.

3.2.2. H5a: Feedback on comment quality is positively associated with changes in feedback quality.

Moving to the next part of Fig. 3 and the third results row of Table 4,

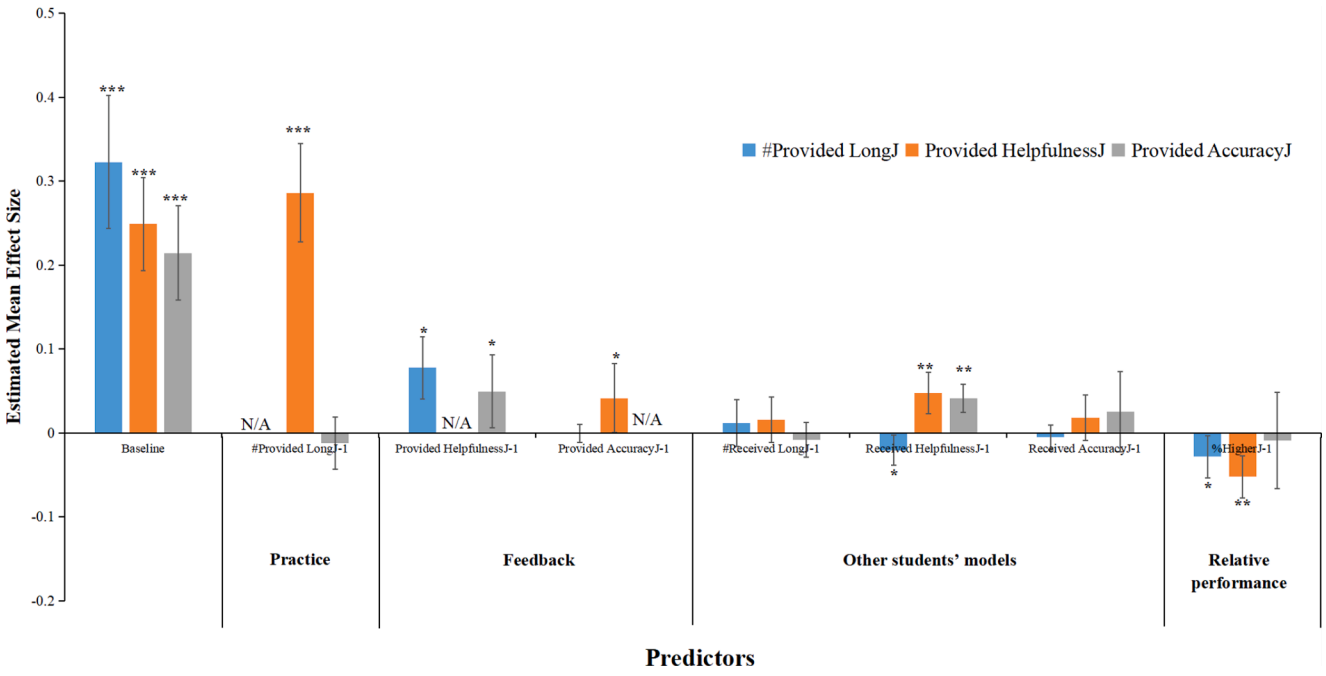


Fig. 3. Estimated mean effect size (with 95% confidence intervals) for each predictor of growth in reviewing quality (*#Provided Long, Provided Helpfulness, and Provided Accuracy).* *=p <.05, **=p <.01, ***=p <.001. N/A indicates the cases in which the predictor is the baseline variable.

receiving more positive feedback on comment quality (Provided Helpfulness$_{J\text{-}1}$) was associated with statistically significant increases that are confidently estimated to be small in both comment length (M = 0.08, 95%CI [0.04, 0.11]) and rating accuracy (M = 0.05, 95%CI [0.01, 0.09]). Thus, H5a was fully supported, albeit with small effect sizes. Interestingly, there was a very large heterogeneity of the effect on length but virtually no meaningful heterogeneity of the effect on rating accuracy across courses and assignments.

3.2.3. H5b: Feedback on rating quality is positively associated with changes in feedback quality.

As shown in Fig. 3 and the fourth results row of Table 4, feedback on rating quality (Provided accuracy$_{J\text{-}1}$) appears to have no general relationship with change in comment length (M = -0.0008, 95%CI [-0.01, 0.01]), alongside no heterogeneity of effect size across courses and assignments; this relationship is ruled out with confidence. By contrast, feedback on rating quality was associated with modest growth in comment helpfulness (M = 0.04, 95%CI [0.0002,0.08]), with a large amount of heterogeneity across courses and assignments. Thus, H5b was partially supported, with suggestions that even the supported case is moderated by contextual factors.

3.2.4. H6: Other students' models of review quality may be positively or negatively associated with changes in review quality.

Fig. 3 and results rows five through seven of Table 4 present the relationship of different aspects of other students' models (*#Received Long*$_{J\text{-}1}$, Received Helpfulness$_{J\text{-}1}$, and Received accuracy$_{J\text{-}1}$) to changes in review quality. Receiving more long comments can be confidently ruled out as a generally important driver of changes in review quality: *#Provided Long*$_J$ M = 0.01, 95%CI [-0.02, 0.04], *Provided Helpfulness*$_J$ M = 0.02, 95%CI [-0.01,0.04], *Provided accuracy*$_J$ M = -0.01, 95%CI [-0.03, 0.01]. However, the first two relationships showed significant heterogeneity, suggesting there may be some situations in which this aspect of models has a small effect.

By contrast, receiving helpful comments was significantly associated with small changes in all aspects of review quality, although negatively with *#Provided Long*$_J$ M = -0.02, 95%CI [-0.04, 0.003], and positively with *Provided Helpfulness*$_J$ M = 0.05, 95%CI [0.02, 0.07] and *Provided accuracy*$_J$ M = 0.04, 95%CI [0.03, 0.06]. Only the negative relationship showed significant contextual heterogeneity.

Finally, receiving ratings in greater accuracy with one another was not significantly associated with any changes in review quality: *#Provided Long*$_J$ M = -0.01, 95%CI [-0.02, 0.01], *Provided Helpfulness*$_J$ M = 0.02, 95%CI [-0.01, 0.05], *Provided accuracy*$_J$ M = 0.03, 95%CI [-0.02, 0.07]. However, there was moderate-to-large heterogeneity in the effect sizes across contexts.

In sum, the comments that students received appeared to matter for their own later reviewing behaviors only with respect to the helpfulness of received comments. Received comment length and rating accuracy did not matter. Thus, H6 received support only with respect to one aspect of received feedback, but it was consistent in the support that aspect received.

3.2.5. H7: Relative performance may be positively or negatively associated with changes in review quality.

As predicted, the rightmost part of Fig. 3 and the bottom row of Table 4 shows that the relative performance of the reviewed document (i.e., reviewing documents of higher quality than one's own) was associated with declines in comment quality (*#Provided Long*$_J$ M = -0.03, 95%CI [-0.04, -0.0004], *Provided Helpfulness*$_J$ M = -0.05, 95%CI [-0.08, -0.03]). However, the effect was small and with a wide estimate in the case of *Provided accuracy*$_J$ M = -0.01, 95%CI [-0.07, 0.05]. Interestingly, there was only significant contextual heterogeneity in the case of number of long comments. Thus, H7 was supported in two of the three possible cases.

4. Discussion

The main aim of the present study was to investigate two primary research questions: 1) the existence of multiple meaningful measures of feedback quality at the level of reviewers within the assignment, and 2) what experiences with peer feedback in the prior assignment in the course were associated with changes over time in each reviewer-level measure of feedback quality. We presented a conceptual framework involving five larger dimensions of feedback quality and then focused on specific conceptualizations of review quality for investigation, focusing

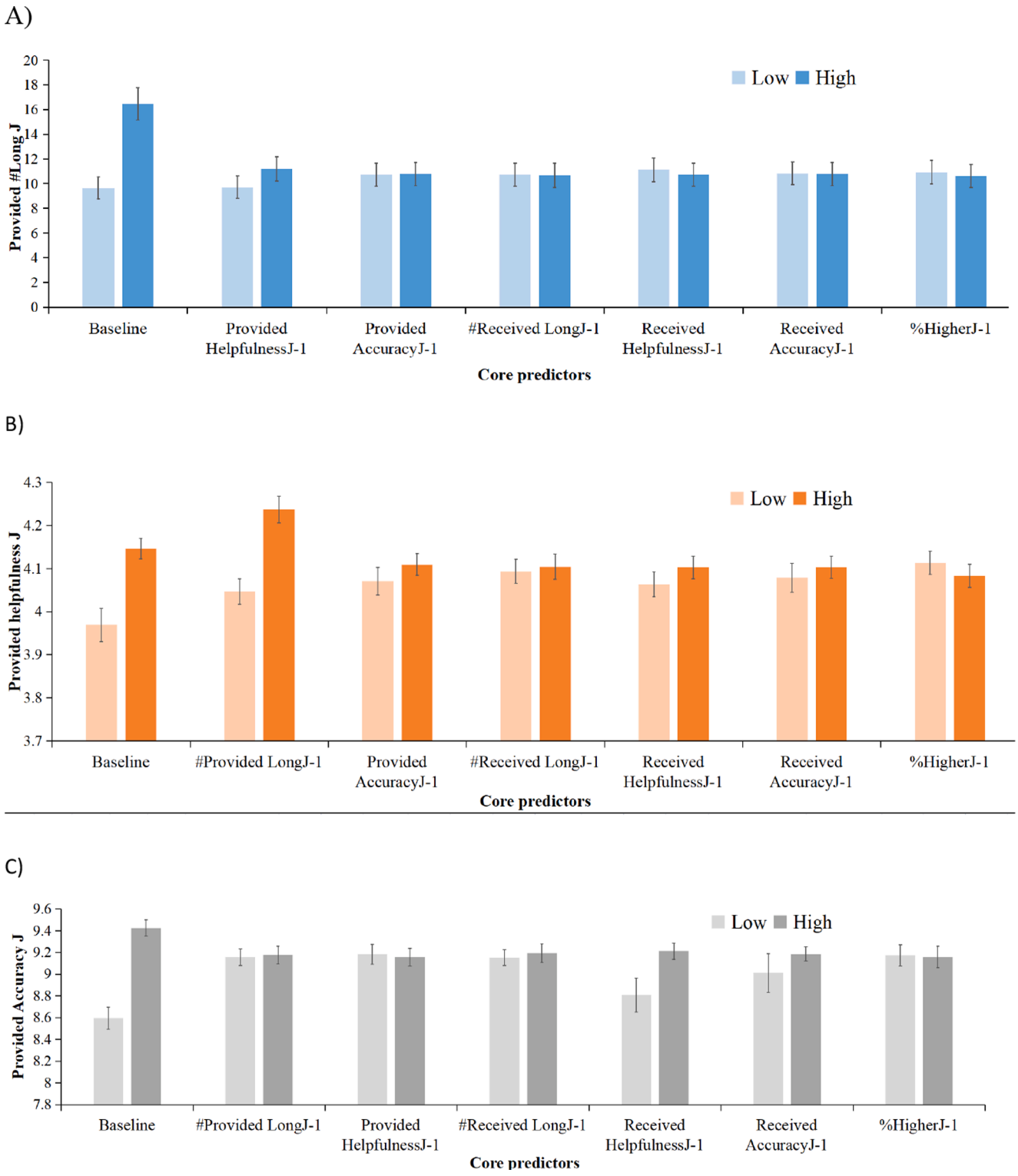


Fig. 4. Estimated marginal means as a function of having high or low levels of a given prior experience on A) provided #Long, B) provided helpfulness, and C) provided accuracy.

on three that could be efficiently investigated in large datasets like the one examined in the current study. We first tested the meaningfulness of these conceptualizations and then explored five hypotheses about the overall relationship between the students' prior experience and their feedback quality in practice, feedback, other' modeling and relative performance. Table 5 summarizes the findings for each of the seven hypotheses.

4.1. The three constructs are meaningfully conceptualized at the level of reviewers within an assignment (RQ 1)

RQ1 was fully supported for all three constructs for all three supporting hypotheses. In particular, comment amount (# long) showed high reliability, and high levels of stability (but not too high), comment accuracy (helpfulness) showed moderate levels of reliability and

Table 5
Degree of support for each hypothesis within each research question.

Hypothesis	Supported?
RQ1: Multiple meaningful measures of feedback quality?	
H1: Reliable	Supported
H2: Semi-stable	Supported
H3: Separable	Supported
RQ2: Are experiences positively or negatively associated with changes in reviewing quality?	
H4: *Practice* (long comments → greater helpfulness, greater rating accuracy)	Partly Supported (helpfulness)
H5a: *Feedback* on comment accuracy (helpfulness feedback → more long comments, greater ratings accuracy)	Supported
H5b: *Feedback* on rating accuracy (rating accuracy feedback → more long comments, greater helpfulness)	Partly Supported (helpfulness)
H6: Other students' *models* (long comments, helpfulness, accuracy → more long comments, greater helpfulness, greater rating accuracy)	Partly Supported (received helpfulness → helpfulness, rating accuracy)
H7: *Relative performance* (reviewed stronger documents → fewer, long comments, lower helpfulness, lower rating accuracy).	Partly Supported (long comments, helpfulness)

moderate levels of stability. Rating accuracy showed moderate levels of reliability and low levels of stability. The two comment quality constructs were moderately correlated with one another and not at all correlated with rating quality. Thus, within the studied contexts, the three review quality constructs were meaningful constructs to examine in quantitative research on peer review literacy (e.g., of their causes and consequences).

Interestingly, we found a nearly zero correlation between comment amount and rating accuracy, raising the question of whether they are both part of a larger construct of peer review quality. Part of the issue may be the relatively lower reliability of the rating accuracy measure: it may be substantially influenced by situational factors like the quality of the document being reviewed. However, we also note that other recent research has found that student's own task performance had little relationship to the accuracy of ratings (Xiong & Schunn, 2021), no relationship at all to being able to accurately detect problems, and only a small relationship to being able to offer helpful advice (Wu & Schunn, 2022). So, reviewing quality might be a domain of relatively separable competencies rather than a domain comprising many highly correlated skills.

4.2. The three constructs showed self-regulated changed in response to reviewing experiences (RQ2)

In terms of the RQ2, the results were mixed. At the highest level, consistent with broader SRL theory (Zimmerman, 2013), every hypothesis regarding changes in reviewing quality received some support; peer review literacy appears to develop through a mixture of learning through practice, feedback, observing others' models, and self-regulating via relative performance. With producing feedback conceptualized as a kind of writing task, these results are then also consistent with other research on the role of SRL in writing (Nückles et al., 2020; Teng, & Zhang, 2018). Further, these mechanisms all appear relevant to self-regulation of reviewing quality, and all four should continue to be explored in future research on different aspects/measures of reviewing quality.

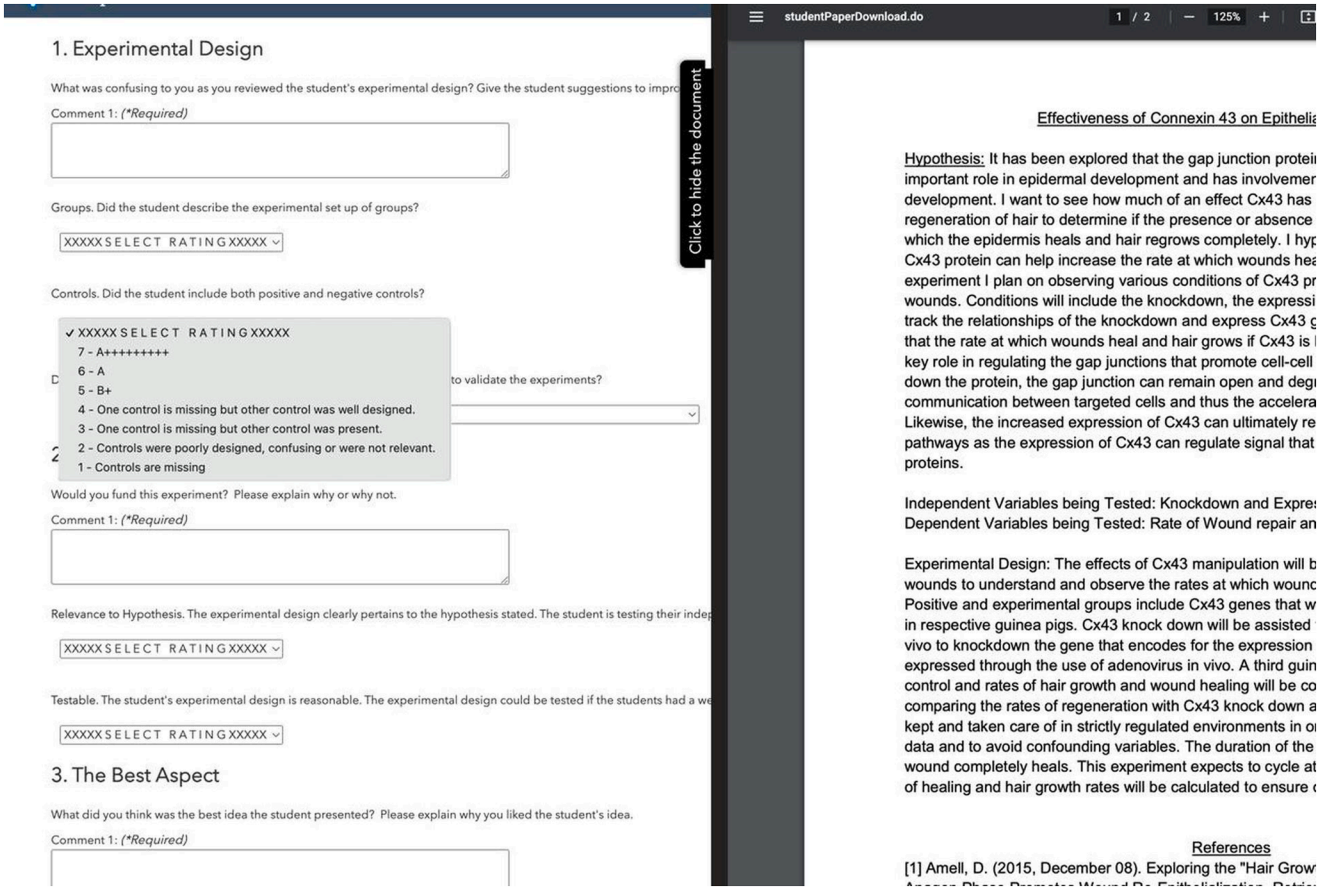


Fig. A1. The screenshot of the peer reviewing interface within Peerceptiv at the time of the study. The reviewed document is available on the right and the reviewing rubric form is on the left. The rubric is divided into reviewing dimensions, which consists of a comment prompt with open-ended text box and one or more rating scales as pull-down menus.

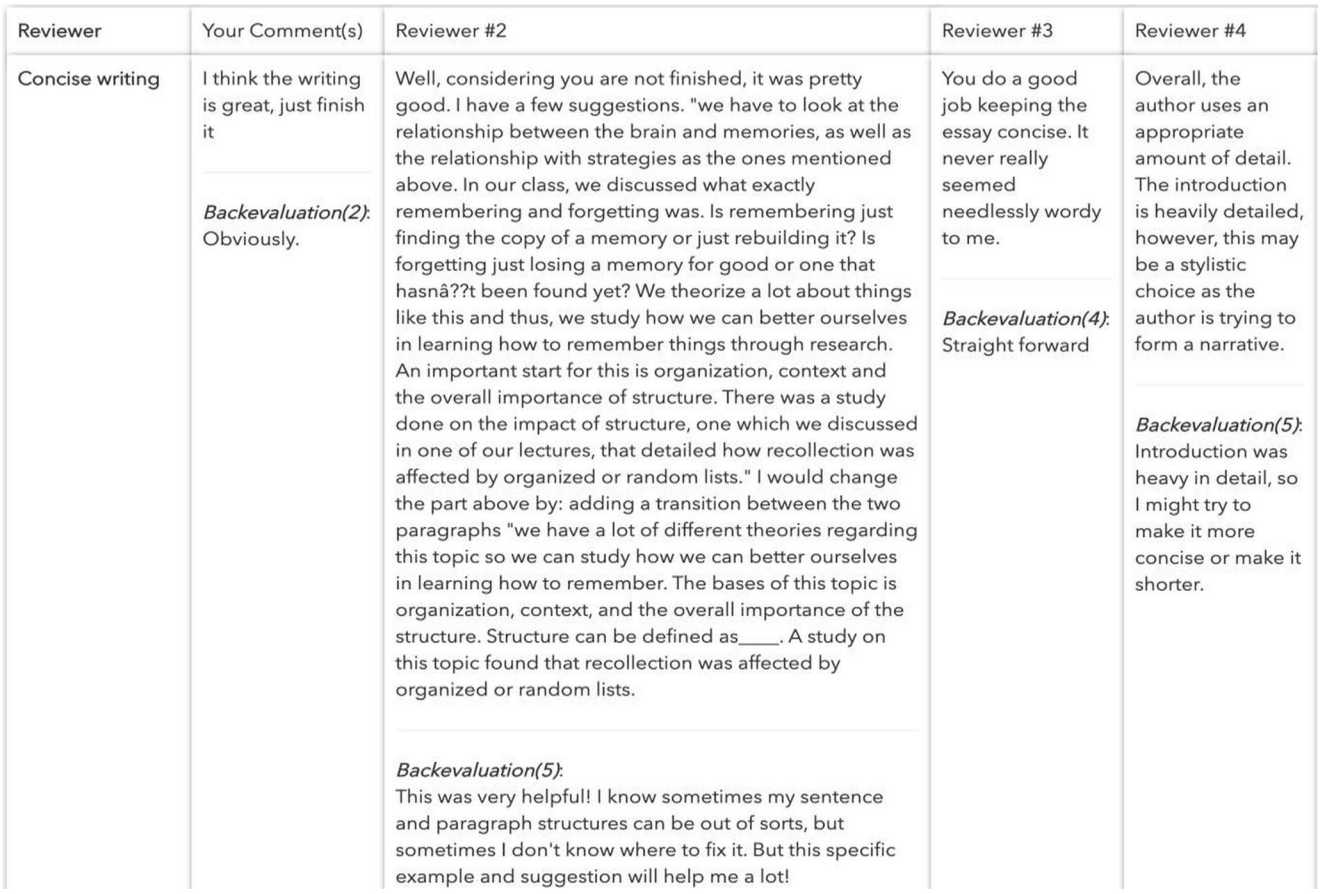


Fig. A2. The Peerceptiv interface at the time of the study showing back evaluations (the entry below each received comment) that one reviewer received (2nd column) for one reviewing dimension (shown in the first column), alongside comments and back-evaluations for comments by other reviewers of the same document (columns 3 through 5).

Table A1
The Cronbach alpha values for each outcome variable based upon contributions of the variable defined separately on each of the first four completed reviews on each assignment.

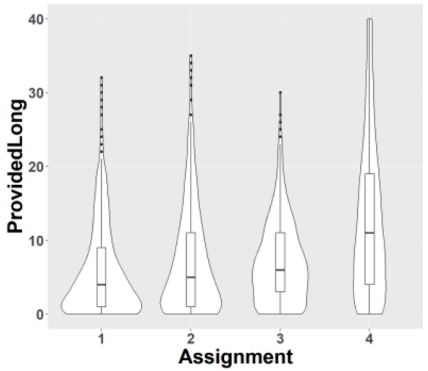
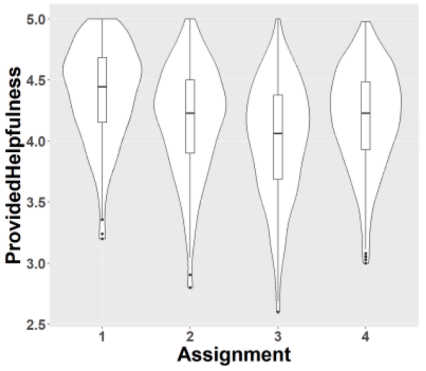
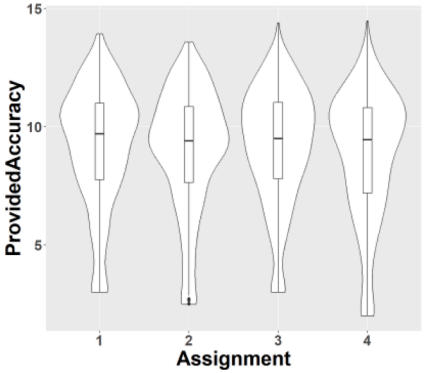
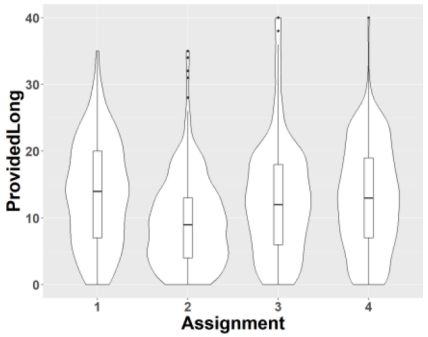
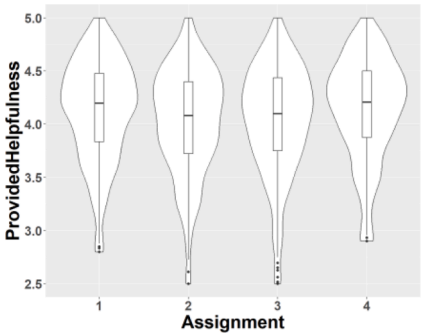
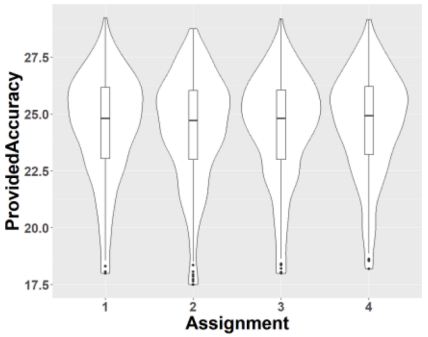
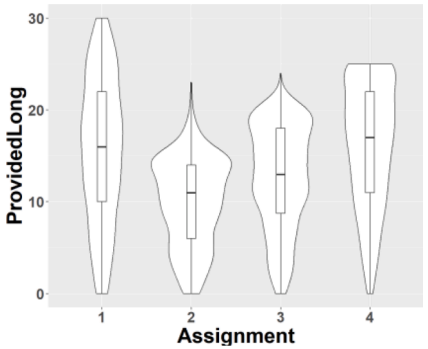
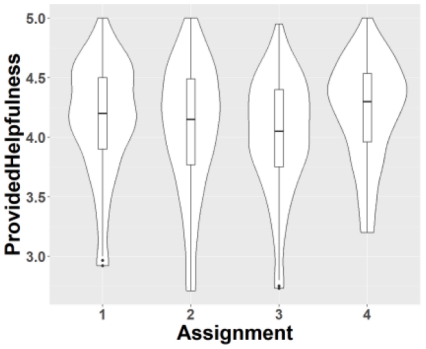
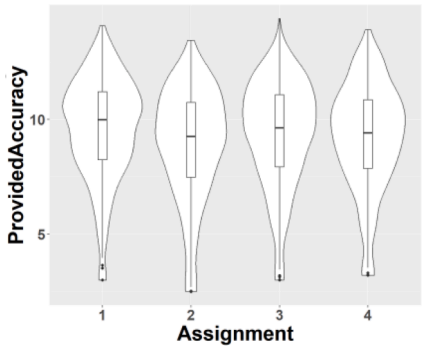
	Biology2015		Biology2016		Biology2018	
	N	alpha	N	alpha	N	alpha
Provided #Long						
1st	621	0.602	640	0.919	726	0.910
2nd	635	0.893	634	0.700	738	0.907
3rd	653	0.870	634	0.967	742	0.916
4th	638	0.860	639	0.922	720	0.924
Provided Helpfulness						
1st	560	0.483	640	0.609	706	0.572
2nd	535	0.475	491	0.575	542	0.613
3rd	601	0.440	637	0.563	644	0.560
4th	576	0.394	520	0.593	666	0.443
Provided accuracy						
1st	560	0.510	545	0.472	706	0.344
2nd	545	0.545	490	0.499	542	0.457
3rd	601	0.466	531	0.500	644	0.433
4th	576	0.561	520	0.450	426	0.459

Turning to the analyses related to RQ2, first, all three constructs/measures showed moderate effect-sizes for the baseline predictor, indicating that they were not so stable as to be uninfluenceable by recent experiences. Second, all three constructs showed some sensitivity to some aspects of prior experiences, lending further support to treating them as meaningful conceptualizations and measures of reviewing quality for investigation in educational research. That is, they appear to be capturing aspects of reviewing quality that is subject to self-regulated change through experience, which is consistent with previous research on SRL (e.g., Raković et al., 2022; Zhang et al., 2022) suggesting that the previous experiences and resulting knowledge shape subsequent behaviors and strategies. However, only changes in helpfulness showed a relationship with all four experience factors. The number of comments produced was only associated with two experience factors (feedback and relative performance), and rating accuracy was also only associated with two experience factors (feedback and models). It could be that these aspects of reviewing quality are indeed only more narrowly regulated, and more subject to situational factors, or it may be that the specific measures did not adequately capture the underlying constructs.

Although not directly the focus of this study, the specific patterns for which experience factors predicted changes in reviewing quality deserve some discussion. Eight of the nine statistically significant relationships were as predicted. In particular, there was support for: more practice in the prior assignment improving comment helpfulness (replicating Zong et al., 2021b); feedback on prior feedback quality improving comment amount (replicating Zong et al., 2022a), comment helpfulness (replicating Patchan et al., 2018), and rating accuracy (replicating Patchan et al., 2018); observing helpful feedback improving comment helpfulness (replicating Zong et al., 2022b) and rating accuracy; and having relative weaker performance leading to declines in comment amount and comment helpfulness. While some of these observed relationships were replications of prior findings, a few were novel empirical findings.

It is also important to acknowledge that half of the tested predictors were not found to be statistically significant, most intensely so for the relationship with the various aspects of other student's models (i.e., only two of nine tested relationships were statistically significant in the predicted direction). The null effects were so clearly clustered around

Table A2
Violin box plot for each variable across assignments in each of the three courses.

	Provided #Long	*Provided Helpfulness*	*Provided accuracy*
Fall2015			
Fall2016			
Fall2018			

other students' models in general and for received length and received accuracy argues against simple noise explanations for the findings overall or the null findings, in particular. Instead, length *per se*, rather than it's correlated variable helpfulness, and received rating accuracy is perhaps not very salient to students. Indeed, in this interface, students are only required to reflect on the helpfulness of the received feedback via the back-evaluation process, and thus they may rarely attend to the other dimensions of the received feedback.

Somewhat surprising in the findings is the unexpected negative relationship between received helpfulness and change in the number of long comments produced. The effect is small, the smallest of all the statistically significant effects, and thus may simply be a spurious finding. However, it is possible that receiving especially helpful feedback is also taken as an implicit criticism of the student's relative skills or that students not acknowledging received critical feedback as helpful are protecting their self-efficacy beliefs. In other case, then, the received feedback maybe active in a similar way as the relative performance information, declining self-efficacy and then leading to reduced engagement in the next assignment (Zong et al., 2021b). However, under this explanation, provided helpfulness should also decline with greater received helpfulness (as it did for relative performance), and it did not. Another possible explanation involves a tradeoff between quality and quantity: as students receive more helpful feedback, they may engage more deeply in providing longer, more helpful comments in the next assignment, but not necessarily as many comments. Further research is needed to replicate and investigate this relationship.

4.3. Limitations and directions for future research

There are three limitations of the current study that warrant discussion. First, although we examined two review quality constructs within the larger conceptual framework (feedback accuracy and amount), we did not examine any measures of the other three main review quality constructs (feedback features, content, and impact). Further, even within the two main review quality constructs that were examined, we did not consider other more specific measures (e.g., comment accuracy as measured by experts). Our strategic focus was on measures that were more readily applied to larger scale quantitative investigations. The same kind of validation methodology could be applied in future studies to other measures of review quality, perhaps leveraging existing datasets that had already had expensive by-hand coding for feedback features and feedback content. NLP could also be applied to automatically detect the comment features (e.g., Darvishi et al., 2022; Nguyen et al., 2014) or comment content (e.g., Ramachandran et al., 2017).

Second, two of the specific measures that were tested had only moderate levels of reliability. The newly proposed measures of rating accuracy (based on rating agreement) had higher reliability than another measure more commonly used in online peer review systems and research (based in rating consistency). However, even higher levels of reliability would be desirable. It is possible that factors influencing growth over assignments in rating accuracy could be revealed with a higher reliability measure. Relatedly, results showed large statistically significant heterogeneity in findings across assignments and courses, indicating there are important situational factors that are moderating the effects. Most saliently, further research will need to investigate whether adjusting for situational factors could improve the reliability of the accuracy of the ratings and comment amount measures. For example, the number of long comments produced by reviewers is likely influenced by the documents they reviewed: high-quality documents have fewer issues to address. Similarly, lower-quality documents are more likely to be evaluated too kindly (Xiong & Schunn, 2021). But other local factors are also likely to matter, like outside distractions, time of day effects, and knowledge of the specific issues in the document being reviewed. Accounting for such document-level effects using more complex algorithms may produce more reliable measures of reviewer-level reviewing quality but will also require additional research to understand the contextual factors.

Lastly, the current study was conducted in the US, at the university level, and in a specific type of biology course. Meta-analyses of different aspects of peer feedback do not typically find substantial educational level or discipline effects on rating validity (Li et al., 2016) or on the learning benefits of engaging with peer feedback (Sanchez et al., 2017). However, the *meta*-regressions revealed some substantial contextual variation within the data studied here. Thus, more meaningful variation in effect sizes is likely when a broader set of contexts is explored, even though the specific effects of education level and discipline may not be significant. Further research should aim to identify the contextual factors influencing the phenomena studied here and the underlying mechanisms driving this variation.

5. Conclusion

The present study investigated the three peer review quality constructs in a large dataset and further explored how the three aspects of peer review quality change through self-regulation experiences. Theoretically, we presented a new five-dimensional conceptual framework for review quality. Then we focused on three constructs that addressed two of the five dimensions, and that could be efficiently investigated in large datasets. Given recent growth in research attention to the concept of peer review literacy as critical for students to develop (e.g., Dong et al., 2023), our findings suggest researchers should not only focus on typical student behaviors but also on the relationships of such behaviors to experience.

Further, we noticed that previous studies on peer feedback tend to focus on one course and often one assignment; based on the significant heterogeneity observed in *meta*-analyses (Li et al., 2016) and within our *meta*-regressions, findings based upon one course and one assignment likely have limited generalizability. We recommend examining variation in effects in addition to variation in psychometric properties (i.e., reliability and discriminant validity) using *meta*-correlation and *meta*-regression at the assignment level. Although recent work has used *meta*-regression across courses, including assignment predictors within the regressions (e.g., Zong et al., 2021a), we found that applying the technique at the assignment level had stronger effect sizes and smaller confidence intervals. Given the growing interest in revealing, understanding, and addressing contextual variation by using learning analytics in enhancing feedback practices (i.e., for whom, under what circumstances, and requiring which supports; Banihashem et al., 2022), this modeling approach is recommended for future studies examining multi-assignment datasets. Furthermore, our results provide empirical evidence supporting SRL as a dynamic and cyclical process that encompassing feedback loops (Lord et al., 2010).

Practically, the findings can lead to some useful suggestions for teachers and educators: 1) Matching authors and reviewers by prior performance review may encourage students to provide more long comments because their long comments appeared to be driven by being recognized for providing helpful comments (which is easier to do for weaker documents) and not having a noticeably weaker document than that of their assessees; 2) comments accuracy appears to improve through practice and therefore scaffolds that strongly encourage longer comments (e.g., guidance on what to include in a comment) will likely improve comment accuracy; 3) rating accuracy appears to be influenced by the accuracy of comments received, and thus efforts focused on generally improving comment accuracy are likely to also lead to improvements in rating accuracy.

Funding

The work was funded by a grant China Scholarship Council [202106770028].

Declaration of Competing Interest

The second author is a co-inventor of the peer review system used in the study.

Data availability

Data will be made available on request.

Appendix A

See Figs. A1 and Fig. A2.
See Table A2.

References

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

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

Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between students' effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review, 32*(4), 979–1002. https://doi.org/10.1007/s10648-020-09569-3

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review™. *Research & Practice in Assessment, 8*, 40–48.

Bandura, A., Grusec, J. E., & Menlove, F. L. (1966). Observational learning as a function of symbolization and incentive set. *Child Development*, 499–506. https://www.jstor.org/stable/1126674.

Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review, 100489*. https://doi.org/10.1016/j.edurev.2022.100489

Benton, T. (2014). Using meta-regression to explore moderating effects in surveys of international achievement. *Practical Assessment, Research & Evaluation, 19*(3), Article n3.

Cambre, J., Klemmer, S., & Kulkarni, C. (2018, April). Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Carlson, P. A., & Berry, F. C. (2008). Using computer-mediated peer review in an engineering design course. *IEEE Transactions on Professional Communication, 51*(3), 264–279. https://doi.org/10.1109/TPC.2008.2001254

Chang, C. Y., Lee, D. C., Tang, K. Y., & Hwang, G. J. (2021). Effect sizes and research directions of peer assessments: From an integrated perspective of meta–analysis and co-citation network. *Computers & Education, 164*, Article 104123. https://doi.org/10.1016/j.compedu.2020.104123

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

Costa, P. J. (2014). Truncated outlier filtering. *Journal of Biopharmaceutical Statistics, 24*(5), 1115–1129. https://doi.org/10.1080/10543406.2014.926366

Cui, Y., Schunn, C. D., & Gai, X. (2021). Peer feedback and teacher feedback: A comparative study of revision effectiveness in writing instruction for EFL learners. *Higher Education Research & Development, 41*(6), 1838–1854. https://doi.org/10.1080/07294360.2021.1969541

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

Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2021a). Learning from reviewing peers' concept maps in an inquiry context: Commenting or grading, which is better? *Studies in Educational Evaluation, 68*, Article 100959. https://doi.org/10.1016/j.stueduc.2020.100959

Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2021b). Giving feedback on peers' concept maps as a learning experience: Does quality of reviewed concept maps matter? *Learning Environments Research, 25*, 823–840. https://doi.org/10.1007/s10984-021-09389-4

Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2022). Does learning from giving feedback depend on the product being reviewed: Concept maps or answers to test questions? *Journal of Science Education and Technology, 31*(2), 166–176. https://doi.org/10.1007/s10956-021-09939-8

Dong, Z., Gao, Y., & Schunn, C. D. (2023). Assessing students' peer feedback literacy in writing: Scale development and validation. *Assessment & Evaluation in Higher Education, Advanced Published Online.*. 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

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. https://doi.org/10.3102/00346543070003

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

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

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

Havnes, A. (2008). Peer-mediated learning beyond the curriculum. *Studies in Higher Education, 33*(2), 193–204. https://doi.org/10.1080/03075070801916344

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). NY: Prentice Hall.

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

Jin, X., Jiang, Q., Xiong, W., Feng, Y., & Zhao, W. (2022). Effects of student engagement in peer feedback on writing performance in higher education. *Interactive Learning Environments. Advanced Published Online.*. https://doi.org/10.1080/10494820.2022.2081209

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

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

Kayapinar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research, 57*, 113–136. https://doi.org/10.14689/ejer.2014.57.2.

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

Kilickaya, F. (2017). Peer assessment of group members in tertiary contexts. In M. Sowa, & J. Krajka (Eds.), *Innovations in languages for specific purposes - Present challenges and future promises* (pp. 329–343). Peter Lang.

Könings, K. D., van Zundert, M., & van Merriënboer, J. J. (2019). Scaffolding peer-assessment skills: Risk of interference with learning domain-specific skills? *Learning and Instruction, 60*, 85–94. https://doi.org/10.1016/j.learninstruc.2018.11.007

Kobayashi, M. (2020). Does anonymity matter? Examining quality of online peer assessment and students' attitudes. *Australasian Journal of Educational Technology, 36*(1), 98–110. https://doi.org/10.14742/ajet.4694.

Latifi, S., Noroozi, O., Hatami, J., & Biemans, H. J. (2021). How does online peer feedback improve argumentative essay writing and learning? *Innovations in Education and Teaching International, 58*(2), 195–206. https://doi.org/10.1080/14703297.2019.1687005

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

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.

Lin, S. S., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17*(4), 420–432. https://doi.org/10.1046/j.0266-4909.2001.00198.x

Lord, R. G., Diefendorff, J. M., Schmidt, A. M., & Hall, R. J. (2010). Self-regulation at work. *Annual Review of Psychology, 61*, 543–568.

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

Margolis, H., & McCabe, P. P. (2006). Improving self-efficacy and motivation: What to do, what to say. *Intervention in School and Clinic, 41*(4), 218–227. https://doi.org/10.1177/10534512060410040401

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

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

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

Nguyen, H., Xiong, W., & Litman, D. (2014, June). Classroom evaluation of a scaffolding intervention for improving peer review localization. In *International Conference on Intelligent Tutoring Systems* (pp. 272-282). Springer, Cham.

Nieminen, J. H., & Carless, D. (2022). Feedback literacy: A critical review of an emerging concept. *Higher Education, 85*, 1381–1400. https://doi.org/10.1007/s10734-022-00895-9

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

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The self-regulation-view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review, 32*, 1089–1126. https://doi.org/10.1007/s10648-020-09541-1

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19*(2), 139–158. https://doi.org/10.1080/10573560308222

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

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

Pastor, D. A., & Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: A tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research, 53*(1), 74–89. https://doi.org/10.1080/00273171.2017.1365684

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

Paré, D. E., & Joordens, S. (2008). Peering into large lectures: Examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning, 24*(6), 526–540. https://doi.org/10.1111/j.1365-2729.2008.00290.x

Peters, L. L., Boter, H., Buskens, E., & Slaets, J. P. (2012). Measurement properties of the Groningen Frailty Indicator in home-dwelling and institutionalized elderly people. *Journal of the American Medical Directors Association, 13*(6), 546–551. https://doi.org/10.1016/j.jamda.2012.04.007

Popken, D. (2020). *The validity and reliability of a single-point rubric to assess student writing performance* (Doctoral dissertation, Western Connecticut State University).

Press, W. H., & Dyson, F. J. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences, 109*(26), 10409–10413. https://doi.org/10.1073/pnas.1206569109

Raković, M., Bernacki, M. L., Greene, J. A., Plumley, R. D., Hogan, K. A., Gates, K. M., & Panter, A. T. (2022). Examining the critical role of evaluation and adaptation in self-regulated learning. *Contemporary Educational Psychology, 68*, Article 102027. https://doi.org/10.1016/j.cedpsych.2021.102027

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

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: A case study. *Journal of Computer Based Instruction, 20*(3), 75–80.

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

Schillings, M., Roebertsen, H., Savelberg, H., van Dijk, A., & Dolmans, D. (2021). Improving the understanding of written peer feedback through face-to-face peer dialogue: Students' perspective. *Higher Education Research & Development, 40*(5), 1100–1116. https://doi.org/10.1080/07294360.2020.1798889

Schunk, D. H., & Greene, J. A. (2017). Historical, contemporary, and future perspectives on self-regulated learning and performance. In *Handbook of Self-regulation of Learning and Performance* (pp. 1-15). Routledge.

Schunk, D. H. (2012). Social cognitive theory. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 101–123). American Psychological Association.

Schunn, C. D. (2016). Writing to learn and learning to write through SWoRD. In S. A. Crossley, & D. S. McNamara (Eds.), *Adaptive Educational Technologies for Literacy Instruction*. Routledge: Taylor & Francis.

Seghier, M. L., Fahim, M. A., & Habak, C. (2019). Educational fMRI: From the lab to the classroom. *Frontiers in Psychology, 10*, 2769. https://doi.org/10.3389/fpsyg.2019.02769

Shabani, K., Khatib, M., & Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Language Teaching, 3*(4), 237–248.

Sherrard, W. R., Raafat, F., & Weaver, R. R. (1994). An empirical study of peer bias in evaluations: Students rating students. *Journal of Education for Business, 70*(1), 43–47. https://doi.org/10.1080/08832323.1994.10117723

Stewart, A. J., & Plotkin, J. B. (2013). From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proceedings of the National Academy of Sciences, 110*(38), 15348–15353. https://doi.org/10.1073/pnas.1306246110

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

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

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*(2), 148–156. https://www.jstor.org/stable/357622.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). error has occurred *International Review of Research in Open and Distributed Learning, 15*(3), 312–327 https://doi.org/10.19173/irrodl.v15i3.1680CopiedAn.

Tan, J. S., & Chen, W. (2022). Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? *Computers & Education, 187*, Article 104467. https://doi.org/10.1016/j.compedu.2022.104467

Teng, L. S., & Zhang, L. J. (2018). Effects of motivational regulation strategies on writing performance: A mediation model of self-regulated learning of writing in English as a second/foreign language. *Metacognition and Learning, 13*, 213–240. https://doi.org/10.1007/s11409-017-9171-4

Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology, 66*, Article 101976. https://doi.org/10.1016/j.cedpsych.2021.101976

Thirakunkovit, S., & Chamcharatsri, B. (2019). A meta-analysis of effectiveness of teacher and peer feedback: Implications for writing instructions and research. *Asian EFL Journal, 21*(1), 140–170.

Thoresen, C. E., & Mahoney, M. J. (1974). *Behavioral self-control*. Holt McDougal.

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.

van den Bos, A. H., & Tan, E. (2019). Effects of anonymity on online peer review in second-language writing. *Computers & Education, 142*, Article 103638. https://doi.org/10.1016/j.compedu.2019.103638

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education, 81*, 123–132. https://doi.org/10.1016/j.compedu.2014.10.001

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

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A, 47*(3), 631–650. https://doi.org/10.1080/14640749408401131

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327. https://doi.org/10.1016/j.learninstruc.2009.08.009

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

Wu, Y., & Schunn, C. D. (2020). 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. (2021a). 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. (2021b). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science, 49*(3), 365–394. https://doi.org/10.1007/s11251-021-09546-5

Wu, Y., & Schunn, C. D. (2022). Assessor writing performance on peer feedback: Exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *Journal of Educational Psychology., 115*(1), 118–142. https://doi.org/10.1037/edu0000768

Xiong, Y., & Schunn, C. D. (2021). Reviewer, essay, and reviewing-process characteristics that predict errors in web-based peer review. *Computers & Education, 166*, Article 104146. https://doi.org/10.1016/j.compedu.2021.104146

Yallop, R. M. A., & Leijen, D. A. (2018). The perceived effectiveness of written peer feedback comments within L2 English academic writing courses. *Eesti Rakenduslingvistika Ühingu Aastaraamat, 14*, 247–271.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., … Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General, 143*(2), 804–824. https://doi.org/10.1037/a0033906

Yuan, A., Luther, K., Krause, M., Vennix, S. I., Dow, S. P., & Hartmann, B. (2016, February). Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1005-1017).

Zhan, Y. (2021). Developing and validating a student feedback literacy scale. *Assessment & Evaluation in Higher Education, 1–14*. https://doi.org/10.1080/02602938.2021.2001430

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

Zhang, Y., Paquette, L., Bosch, N., Ocumpaugh, J., Biswas, G., Hutt, S., & Baker, R. S. (2022). The evolution of metacognitive strategy use in an open-ended learning environment: Do prior domain knowledge and motivation play a role? *Contemporary Educational Psychology, 69*, Article 102064. https://doi.org/10.1016/j.cedpsych.2022.102064

Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology-facilitated peer assessment and supporting strategies: A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(3), 372–386. https://doi.org/10.1080/02602938.2019.1644603

Zhu, Q., & Carless, D. (2018). Dialogue within peer feedback processes: Clarification and negotiation of meaning. *Higher Education Research & Development, 37*(4), 883–897. https://doi.org/10.1080/07294360.2018.1446417

Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist, 48*(3), 135–147. https://doi.org/10.1080/00461520.2013.794676

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance*. Routledge/Taylor & Francis Group.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. https://doi.org/10.1037/0022-0663.81.3.329

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*, Article 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 & Evaluation in Higher Education, 46*(6), 973–992. https://doi.org/10.1080/02602938.2020.1833179

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

Zong, Z., Schunn, C. D., & Wang, Y. (2022b). Do experiences of interactional inequality predict lower depth of future student participation in peer review? *Computers in Human Behavior, 127*, 107056. https://doi.org/10.1016/j.chb.2021.107056

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