Computers in Human Behavior 124 (2021) 106924

Contents lists available at ScienceDirect

Computers in Human Behavior

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

What aspects of online peer feedback robustly predict growth in students' task performance?

Zheng Zong [a], Christian D. Schunn [b,*], Yanqing Wang [a]

[a] School of Management, Harbin Institute of Technology, Harbin, 150001, China
[b] Learning Research & Development Center, University of Pittsburgh, PA, 15260, USA

ARTICLE INFO

Keywords:
Peer assessment
Peer feedback
Students' performance
Feedback amount
Feedback length
Course context

ABSTRACT

The value of online peer feedback in education has been widely established, and the use of online peer feedback tools is rapidly growing in practice. However, effect sizes appear to vary widely across studies, suggesting implementation details matter substantially. Further, there remain open questions about exactly which aspects of the multi-faceted peer feedback experience are most closely associated with learning outcomes. Within 13 different courses involving 2421 students, drawn across seven universities and six content disciplines, temporally lagged multiple-regression analyses were used to test the unique contributions of quantity, depth, and quality of received and provided comments to students' growth in task performance across assignments. Meta-regression is applied to precisely estimating overall effect sizes and variation in effect sizes. Results reveal stronger relationships with growth in task performance for 1) provided rather than received comments, 2) longer rather than more comments, and 3) comments perceived to be helpful for revision. Further, there was largely quantitative and qualitative variation in observed relationships across courses that not attributable to statistical noise.

1. Introduction

Peer review, combining quantitative assessment and qualitative feedback aspects, is a rich student-centric learning technique that can be broadly implemented into essentially any kind of course from elementary to advanced tertiary education (Applebee & Langer, 2011; Min, 2016). It is particularly useful in combination with performance tasks like essays, presentations, or projects (Bijami et al., 2013). Further, it is a kind of learning technique that has broadly grown immensely in popularity (Double et al., 2019), especially through the affordances offered by computer-based, online peer review (Lu & Law, 2012; Zheng et al., 2020). Online peer review facilitates asynchronous work (submitting and reviewing), management of document distribution to reviewers, double-blind reviewing, smart scaffolds for effective reviewing, efficient grading of documents and reviewing behaviors, and calculation of a variety of rich metrics about individual-level and assignment level performance (Cho & Schunn, 2007; Topping, 1998; Zheng et al., 2020).

Studies on various aspects of peer review in instruction have exploded (Double et al., 2019), resulting in several recent systematic reviews (van Popta et al., 2017; Panadero & Alqassab, 2019) and meta-analyses (e.g., Wright & Jenkins-Guarnieri, 2012; Huisman et al., 2019; Double et al., 2019; Li et al., 2020), including a meta-analysis focused on technology-supported peer review (Zheng et al., 2020). The reliability and validity of quantitative peer assessments are generally as good as those produced by teachers (Wright & Jenkins-Guarnieri, 2012; Li et al., 2016), particularly when properly scaffolded (Cho & Schunn, 2007) and when averaging across multiple peers (Marsh et al., 2008). Further, peer review has been broadly shown to improve student learning outcomes (Wright & Jenkins-Guarnieri, 2012; Double et al., 2019; Huisman et al., 2019; Li et al., 2020), as well as supporting social-affective development and metacognitive awareness (Sluijsmans et al., 2001; van Gennip et al., 2010).

However, the average learning benefit is modest, and there is considerable heterogeneity of learning effects (Huisman et al., 2019). Some of the heterogeneity of effect is explained by the context in which the peer review occurs (e.g., discipline, grade band, type of learning objects), but they account for relatively small amounts of variance (Huisman et al., 2019; Langan et al., 2019; Li et al., 2020). Exploration of other potential factors is needed. Peer review has many steps/components, most notably providing feedback vs. receiving peer feedback, and there is a wide variety of ways each step can be implemented and scaffolded. It is likely that variation in these steps plays an

* Corresponding author. Learning Research & Development Center, University of Pittsburgh, Pittsburgh, USA.
E-mail address: schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.chb.2021.106924
Received 8 March 2021; Received in revised form 10 June 2021; Accepted 14 June 2021
Available online 18 June 2021

important role in shaping learning opportunities for students. This is the focus of the current investigation: how do the different aspects of peer-reviewing shape how much students learn?

1.1. Theoretical background

Peer review has generally been justified in terms of learning from feedback (Hattie & Timperley, 2007), learning in the Zone of Proximal Development (Shabani et al., 2010), self-regulated learning (Winne, 2010), and learning via constructive or interactive processes (Chi & Wylie, 2014). More specific theoretical frameworks have also been developed to deconstruct peer review at various grain sizes (e.g., Topping, 1998; Narciss, 2008; Strijbos & Sluijsmans, 2010; Gielen & De Wever, 2015). In terms of the quantitative evaluation vs. qualitative commenting aspects of peer review, it appears that the qualitative commenting aspect is most important for supporting student learning (Wooley et al., 2008; Huisman et al., 2019), likely because qualitative commenting contains more information about how to improve for the feedback recipient and requires more constructive activity by the feedback provider.

1.2. Providing vs. receiving peer feedback

In terms of learning outcomes, most studies have considered only the effect of receiving peer feedback (e.g., Cho & Schunn, 2007; van Popta et al., 2017; Tsivitanidou et al., 2018; Wichmann et al., 2018) or considered only the combination of receiving and providing peer feedback as a whole (Lundstrom & Baker, 2009). However, a few studies have considered the relative contributions of providing vs. receiving peer feedback, and the findings generally support a role for both but a relatively larger role of providing feedback than receiving feedback in terms of perceptions of learning outcomes (e.g., Ion et al., 2019; Schunn, Godley, & DeMartino, 2016) or actually changes in student performance (e.g., Cho & MacArthur, 2011; Lundstrom & Baker, 2009; Traga Philippakos et al., 2018; Wu & Schunn, 2021). However, a systematic contrast of the relative impact of providing vs. receiving feedback has not yet been conducted, particularly in naturalistic contexts in which both feedback providing and receiving co-occur (Wu & Schunn, 2021) and potentially interact to influence learning outcomes (Cho & Schunn, 2017).

In terms of underlying mechanisms for the benefit of receiving peer feedback, any form of feedback that has been received can provide an important learning opportunity (Gielen et al., 2010), particularly when students act on the feedback they receive (Wu & Schunn, 2021). In terms of peer feedback vs. self-assessment or teacher feedback, students can receive deep knowledge from their peers to address their own knowledge gaps (Vickerman, 2009; Davey, 2011), although inaccuracies in the feedback received from peers may limit learning benefits (Nelson & Murphy, 1992). In addition, peers can detect issues more readily in their peers' documents than in their own documents (Baturay, 2015; Roscoe & Chi, 2007). Further, feedback from peers may be conveyed in a particularly understandable way or at the right level of explanation (Butler et al., 2013), although peers can also perceive the feedback they receive from peers as less expert and therefore resist it (Kaufman & Schunn, 2011; Nelson & Murphy, 1992). Finally, the greater volume of received feedback from multiple peers (relative to a single instructor) may be especially productive for revision and learning (Cho & Schunn, 2007; Wu & Schunn, 2020b; Zhang, Schunn, & Baikadi, 2017; Zou et al., 2018). However, students can also become de-motivated and disengage from further learning tasks if they receive too much negative feedback (Scott et al., 2019).

In terms of underlying mechanisms for providing peer feedback, providing feedback to peers provides an opportunity to learn from mistakes peers also make as well as from seeing good models (Patchan et al., 2016). In addition, students may improve their ability to detect and address problems through practice on their peers' documents because they are more likely to detect problems in their peers' documents (Baturay, 2015). Further, providing feedback is both more active a format of learning activity (than just receiving feedback) and can contain elements of constructive learning (van Popta et al., 2017), which is particularly helpful for learning (Chi & Wylie, 2014). However, peers may have different problems, especially across peers at different performance levels (Patchan et al., 2016), and thus the practice of revising their peers' documents may not be efficient in terms of focusing on issues within their zone of proximal learning (Wu & Schunn, 2021). Finally, in a naturalistic context, time is a limited quantity. As students spend more time providing feedback to their peers, they may need to spend less time working on revising their own documents.

In sum, there are a wide variety of potential benefits, but also potential limitations, of both providing and receiving peer feedback. For these reasons, more research is needed that systematically examines the relative learning benefits of each. This gap is important for theoretical models of learning from peer review. It is also important from a practical perspective in terms of optimizing peer review arrangements for learning outcomes. For example, if most of the benefit is from providing, systems could be set up in which students only provide feedback to fictional students (e.g., as in Lundstrom & Baker, 2009), potentially replacing the need to produce documents, organize assessor-assessee pairs, wait for feedback, etc.

1.3. Unpacking Amount, Depth, and Quality of Peer Feedback

Many of the studies considered in the prior section have focused on the *amount* of comments provided and received. For example, Cho and MacArthur (2011) found that students who provided more total comments showed greater learning outcomes. Similarly, Wu and Schunn (2021) showed that students who received more comments made more revisions to their documents and then improved to a greater extent on the next assignment. Further, Zou et al. (2018) found that students were more likely to act on a comment when multiple peers addressed the same issue, and logically that is more likely to occur if each student is required to provide many comments. But what about variation within a given comment? Comments that peers give each can vary from a couple of words (e.g., "Good job!") to a very long paragraph.

While the quantitative aspect of peer assessment is relatively well structured, by its very nature, peer comments are inherently more open-ended. It is therefore not surprising that a wide variety of frameworks have been proposed to characterize variations in the comments that peers provide and receive from each other. For example, peer feedback has been divided into evaluation vs. information (Narciss, 2008), simple vs. detailed/elaborated (Narciss, 2008; Strijbos & Sluijsmans, 2010), feedback with varying degrees of structure (Gielen & De Wever, 2015), and feedback varying in terms of features that it contains such as praise/criticism, problems vs. solutions, general vs. specific solutions, and descriptions vs. explanations (e.g., Cho & MacArthur, 2011; Nelson & Schunn, 2009; Tseng & Tsai, 2007; Wu & Schunn, 2020a). Collectively, we refer to these distinctions as ones of comment *depth* in that some comment forms are considered deeper than others, but the variations will likely be correlated with one another and reasonably approximated by comment length (e.g., Patchan et al., 2018). Most of the research on various aspects of comment depth has focused on the receiving side (i.e., which features of received feedback matter) and on its benefit for document revision rather than tracking learning effects (i.e., whether future performance on new tasks is improved). For example, receiving elaborated feedback or feedback with problems, solutions, and explanations results in more document revisions (Gielen & De Wever, 2015; Nelson & Schunn, 2009; Patchan et al., 2016; Strijbos & Sluijsmans, 2010; Wu & Schunn, 2020a). It is not yet known which of these aspects of comment depth best support student learning (rather than only revision), although it is known that more revision behaviors in response to peer feedback are related to improvements in student learning (Wu & Schunn, 2021).

Pragmatically speaking, instructors (in setting up peer review

assignments) and students (in completing peer review assignments) encounter an amount vs. depth tradeoff. For a given amount of time, a student might comment on more issues, or they might comment on a smaller number of issues in greater depth. Instructors can require a minimum number of comments on different aspects of an assignment, or they can require students a minimum word length for each comment (Patchan et al., 2018). Prior research has not considered the relative benefits of amount vs. depth. Is it better to ask students to comment on more issues, or is it better to ask students to provide more elaborated/deeper/longer comments? This tradeoff might play itself out differently from the providing and receiving sides of peer review.

Another way that researchers have conceptualized variation in peer comments is in terms of perceived quality. A given comment might be perceived as not understandable, not actionable, incorrect, insufficient, or not persuasive, and these dimensions have been associated with the likelihood of acting on peer feedback (e.g., Nelson & Schunn, 2009; Kaufman & Schunn, 2011; Patchan et al., 2016; Huisman et al., 2018; Wu & Schunn, 2020a). A number of online peer review systems for instruction now include a step in which the recipients of feedback are asked to judge the quality/helpfulness of the feedback that they received (e.g., Cho & Schunn, 2007; Patchan et al., 2016; Misiejuk et al., 2020). Giving students a grade incentive based upon these ratings appears to improve feedback length (Patchan et al., 2018). However, the relationship of perceived comment helpfulness to learning outcomes has not been investigated. Students might overvalue easy-to-act-upon comments, which may not be as productive for learning (Chi & Wylie, 2014). Further, comments that are especially helpful to the recipient might not be especially helpful to the provider's learning. For example, a comment might be unimportant for the assessee's document but could be more important for the assessor's learning concerns.

In sum, there are many open questions about what kind of peer feedback experiences are particularly productive for student learning, both on the receiving side and on the providing side. Comment amount, comment depth, and perceived comment quality are broadly applicable constructs but have not been systematically investigated in terms of their relationship to student learning outcomes, especially in relationship to one another.

1.4. Contextual variation in peer feedback benefits

There has been growing emphasis on the role of context in shaping educational outcomes overall (Curran, Gustafson, et al., 2019a, 2019b), and, in particular, support, hindering, and otherwise shaping the efficacy of different learning methodologies. Learners, classrooms, and schools all vary in a wide variety of ways, and the science of instruction must begin to understand the nuances of for whom and under what circumstances of educational findings and educational interventions. Online peer feedback is now broadly implemented across educational levels and course disciplines (Huisman et al., 2019; Li et al., 2020; Wright & Jenkins-Guarnieri, 2012). On the one hand, the findings of research on online peer feedback now have very broad applicability, no longer being limited to rarely used or fragile instructional technologies. On the other hand, expectations should be growing that research attend to the wide variety of course types in which online peer review is used. To match this expectation, the spread of particular online peer feedback platforms creates opportunities for larger-scale investigations of patterns across courses, disciplines, and universities (e.g., Leijen & Leontjeva, 2012; Misiejuk et al., 2020; Ramachandran et al., 2017).

Today, few studies have formally examined how the contexts of peer assessment might shape the benefits obtained from peer feedback. As noted earlier, meta-analyses have document substantial variation in learning benefits of peer feedback but also found that simple predictors like course level or discipline accounted for little of the variance. It is also important to note that one fundamental challenge to meta-analyses, which normally look across different studies, is the need to ignore potentially important variation in measurement methods. Each author team will have instantiated both independent and dependent variables in slightly different ways, and it is unclear how much of the effect heterogeneity is due to this measurement variation. In the case of online peer feedback, different systems implement the peer review process in slightly different ways, which not only can influence patterns but also can shape how data is processed (e.g., how separate comments are counted). One approach to addressing this challenge is the use of a common methodological approach to a large dataset of many courses all using a shared underlying online peer review technology. If considerable effect heterogeneity still exists in such a dataset, it must be meaningful variation in the benefits of peer review processes rather than due to measurement variation or different technological mediation.

Based on these critical gaps in the literature on peer feedback, this study explores major open questions about the nature of peer reviewing experiences that influence growth in students' performance. It leverages a large dataset involving naturally collected data (i.e., not manipulated or otherwise setup for experimental purposes) from large courses from a variety of disciplines and universities all using the same online review platform. It applies a regression approach to estimate the strength of relationships between different aspects of reviewing and growth in student performance across assignments. In particular, the study asks four key research questions.

RQ1: Are gains in task performance most closely associated with the amount of comments or comment depth (for received or provided comments)?

RQ2: Are student perceptions of comment value associated with gains in task performance (for receivers or providers of comments)?

RQ3: Are gains in task performance more consistently associated with (quantitative dimensions of) receiving comments or providing comments?

RQ4: Which of these relationships vary substantially across courses (while holding the underlying technology constant)?

Based upon a few isolated studies in the literature, we expected larger effects associated with comment depth than with amount of comments (RQ1) and with providing than with receiving (RQ3), but without specific predictions for relative effect sizes within received or providing comments. RQ2 and RQ4 are open research questions without *a priori* hypotheses.

2. Materials and methods

2.1. Course settings and participants

Participants were 2421 university students who were enrolled in one of 13 undergraduate courses across seven universities, distributed broadly across the US, representing moderately selective to very selective, large public universities. The courses were from five different disciplines (Astronomy, Business, Biology, Entomology, Psychology, and cross-sciences laboratory course) selected to span larger discipline categories (Business, Natural Sciences, Social Sciences) that are often treated separately in meta-analyses in general (e.g., Schneider & Preckel, 2017; Sisk et al., 2018; Wright & Jenkins-Guarnieri, 2012) as well as in the meta-analyses of peer feedback. Including a range of disciplines can thereby improve the generalizability of the findings. These courses all used a shared online peer assessment system which facilitated data collection and consistent variable definition. Further, these courses all implemented peer assessment for at least four different assignments (maximum of seven assignments), which to support longitudinal analyses of task performance gains across assignments.

The courses varied in enrolment size, but the mean and median enrolment was close to 200 students, and only one course had enrolments well below 100 students. Predominantly large courses were selected to increase statistical power and because most of the feedback students in these courses received on their submissions was likely to be from peers rather than instructors or TAs. All but three of the courses (Biology 2, 3, and 4) were introductory courses, likely as a result of the

focus on larger enrolment courses.

The student demographics varied widely across courses (see Table 1) by age, from a mean age of 19 to a mean age of 21 (which likely also closely mirrors how long students in each course had already attended the university), gender, from predominantly female to majority male, and by race/ethnicity (from predominantly White to predominantly Asian). These variations in demographics are likely a function of both university and discipline base-rates. Although not a focus in the analyses, this contextual variation further supports the generalizability of findings.

3. Materials

Document submission, peer reviewing, comment quality evaluation, and demographic data collection were all managed by the widely-used online peer reviewing system called *SWoRD/Peerceptiv* (Cho & Schunn, 2007; Patchan et al., 2016). The system also provided access to data through downloads organized by a system ID to allow for linking data sources without violating student privacy. The data for analysis was obtained from the server in an anonymized format. Analysis of anonymized data was approved by the University of Pittsburgh Institutional Review Board.

All peer reviewing was double-blind, with authors identified by a pseudonym and reviewers by a number. The system converted students' document submissions into a PDF format for reviewing, and students could view the PDF while completing the reviewing form, which involved a mixture of commenting and rating prompts (see Fig. 1). The instructors organized the review into one or more dimensions, and each dimension had one comment prompt (with an instructor selected minimum and maximum number of textboxes) and one or more rating dimensions, usually on a 1–7 scale with instructor-provided anchor descriptions for each rating level. Each student had to complete between three and six reviews (determined by the instructor). Although the system allows the submission of group-based assignments, all of the assignments in this dataset involved individual-based assignments. Reviewing details such as the contents and number of reviewing dimensions could and usually did vary from one assignment to the next within a course.

The system gave students grades for reviewing accuracy, based upon the correlation between the ratings a student gave across rating dimensions and reviewed documents and the mean ratings produced by others in the course who reviewed the same documents. The system also gave students grades for comment helpfulness, based on helpfulness ratings (on a 1–5 Likert scale) each author gave their reviewers after receiving their comments. Students also receive a task grade, which is simply points for completing the reviewing and comment helpfulness rating tasks. These grading incentives improve the consistency of participation in the peer reviewing process (Patchan et al., 2018). The system also emailed students reminders, when needed, of upcoming submission and reviewing deadlines to further encourage student participation. The ratings are the source of task performance measures, and the comments are the source of the received/provided (number of textboxes used and length of comments in each textbox).

3.1. Measures

All measures were derived from data automatically collected within *Peerceptiv*. To prepare the data for analysis, the data were organized for each student on each assignment (e.g., overall task score on each assignment, number of comments received on that assignment, number of comments provided on that assignment). Table 2 presents a summary of the variables. Mean values for each variable within each course are presented in the Appendix (Table A1).

Task Performance. Performance in each assignment was assessed through the peer review process, and these ratings are used to produce a measure of task performance. A mean across rating dimensions and reviewers is used to produce a score. To correct for differences in task difficulty and changes in rating rubrics across assignments, a standardized score was created for each assignment (i.e., subtracting out the assignment-specific mean and dividing by the assignment-specific standard deviation), called *Task Score*$_J$, representing the relative task performance on the J^{th} assignment. For the purposes of the multiple regression analyses used here, it does not matter whether the students who receive or provide more reviews are improving their task performance in absolute terms or only in relative-to-peers terms.

A recent meta-analysis of a large number of peer assessment studies generally found high reliability and validity of ratings produced by peers, with generally small moderation by contextual features (Li et al., 2016). The *Peerceptiv* system embeds a number of supports that improve the reliability and validity of the ratings, such as averaging across dimensions/reviewers, concrete anchors for each rating level, entering comments before ratings, and grading incentives for producing accurate ratings and helpful comments. Thus, it is not surprising that a study examining the validity of mean student ratings in the *Peerceptiv* system found very high correlations between mean student ratings and both teacher ratings and trained expert ratings (Patchan et al., 2016). Since a different random set of peers evaluate documents in each assignment, there is also no embedded confound between comment providing in one round and relative task performance growth observed into the next round.

Amount of comments provided and received. While participants were required to evaluate their peers' work for a fixed number of peers (e.g.,

Table 1
For each course, course discipline, university code, # of participating students, mean age, % female, and % reporting each race/ethnicity.

Course	University	# students	Mean age	% female	Race/Ethnicity			
					% Asian	% Black	% Latinx	% White
Astronomy	A	277	21	45%	4%	1%	–	95%
Biology 1	B	98	19	82%	42%	3%	28%	28%
Biology 2	C	196	20	60%	76%	4%	–	20%
Biology 3	C	274	20	59%	69%	2%	14%	15%
Biology 4	C	209	21	59%	69%	1%	18%	12%
Business	D	182	20	49%	6%	3%	5%	86%
Entomology 1	E	296	–	–	–	–	–	–
Entomology 2	E	196	–	–	–	–	–	–
Entomology 3	E	185	–	–	–	–	–	–
Entomology 4	E	103	–	–	–	–	–	–
Psychology 1	F	194	–	–	–	–	–	–
Psychology 2	F	166	21	63%	8%	10%	–	82%
Laboratory science	G	45	19	76%	20%	10%	3%	67%

Note. Demographics details are self-reported by students. - = not recorded. Some courses directly linked the peer review system with the Learning Management system, and the demographic self-reporting step is skipped in those cases. Three courses were from an early instantiation of the system that did not include a Hispanic/Latinx reporting option due to a programming error.

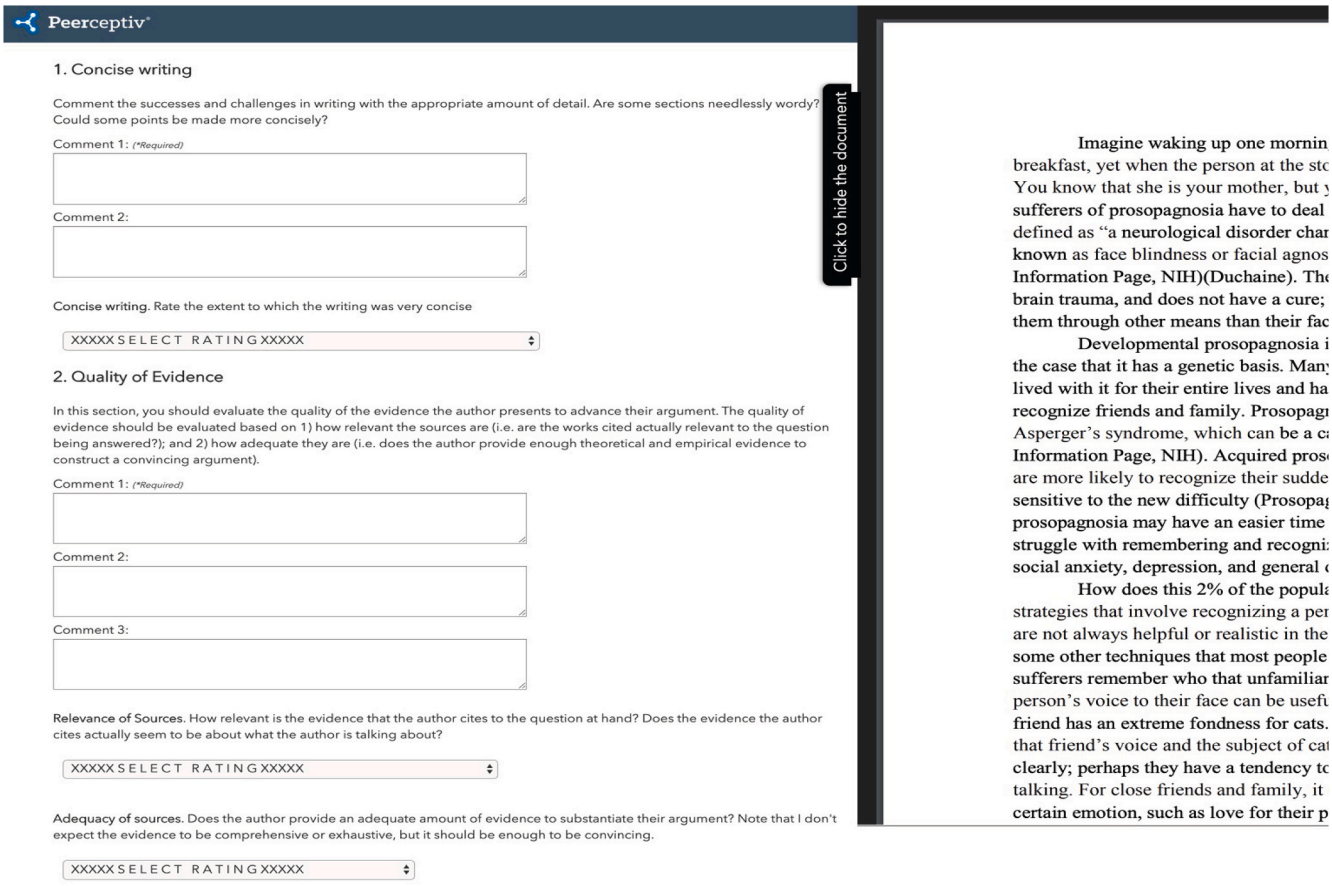

Fig. 1. The reviewing interface in *Peerceptiv* from the time of the study. On the right is a PDF document viewer. On the left, students enter comments for each dimension in the available textboxes and make ratings using pull-down menus. The rating form details are selected/created by the instructor and often vary by assignment.

Table 2
Constructs, specific measures, and the definitions.

Constructs	Measure	Definition
Task Performance	Task Score$_J$	The standardized score of the student's task on the Jth assignment (i.e., score minus mean assignment grade/SD of the assignment grade)
Amount of comments provided	Amount Provided$_J$	The number of peer comments provided by a student across all reviews completed for the Jth assignment.
Length of comments provided	Length Provided$_J$	The total number of words provided by a student on the Jth assignment.
Helpfulness of comments provided	Helpful Provided$_J$	The mean helpfulness rating for provided comments on the Jth assignment.
Amount of comments received	Amount Received$_J$	The number of peer comments received by a student across all received reviews on the Jth assignment.
Length of comments received	Length Received$_J$	The total number of words received by a student on the Jth assignment.
Helpfulness of comments received	Helpful Received$_J$	The mean helpfulness rating for received peer comments on the Jth assignment.
Assignment number	J	The sequential peer review assignment number within a course

3,4, or 5, depending on the course), sometimes students did not complete all the required reviews, and sometimes they completed bonus reviews. Moreover, instructors usually setup the reviewing form with a required single comment textbox for each reviewing dimension and the option to provide two or three comment textboxes in each dimension (as shown in Fig. 1). Empirically, the number of comments included in a review varies somewhat by the document quality, a greater amount by the reviewer tendencies, but a large variation is simply between different reviews for a given document (Patchan et al., 2016). Therefore, collectively across reviewing dimensions and reviews completed for a given assignment, there could be wide variation in how many comments a student received on their submission for assignment *J* (*Amount Received$_J$*) or how many comments they provided to their peers on that assignment (*Amount Provided$_J$*), and these amounts are not highly confounded with author or reviewer characteristics. The value for the amount received was recorded as missing when no document was submitted for a given assignment (mean %missing = 1%).

This approach treats the number of textboxes used as a meaningful unit (i.e., roughly representing the number of issues examined/discussed). Prior research has successfully connected the number of comments provided to learning outcomes using this automatic counting approach (Cho & MacArthur, 2011). Sometimes students do include multiple idea units within one comment box. However, when comments collected using the *Peerceptiv* interface have been carefully segmented for separate idea units (i.e., one textbox comment that actually discusses two different issues), such additional segmentation is needed in only a minority of cases (i.e., less than 10% of comments).

Across reviews for a single document, there could be some redundancy in the content of the comments, and this approach treats each comment as a separate contribution, regardless of such redundancy. Similarly, a given student might give the same comment to multiple authors in a given assignment. However, redundancy, when formally coded, turns out to be rare (each reviewer finds a different subset of the issues to be addressed; Wu & Schunn, 2020b). Further, when

redundancy does happen, it tends to be beneficial (e.g., much more likely to be addressed in a revision; Wu & Schunn, 2020b).

Length of comments provided and received. The *Peerceptiv* interface has no required minimum length for comment; a single word entered into the textbox would be accepted by the system (although likely to generate a low helpfulness rating). Students could also provide a whole paragraph in the textbox, carefully describing the nature of an issue in the document, the location(s) of the issue, an explanation for why the issue is problematic, and a suggestion for how to address the issue (Patchan et al., 2018; Wu & Schunn, 2020a). For example, here are two individual comments sampled from one assignment in one of the courses in the dataset:

1) The paper was weak in comparing the inductive and hypothetico-deductive methods against each other. While there was one or two comparisons, some of the main points were missing. By including how induction doesn't allow for imagination and logical fallacies of induction would help.
2) The topic is too simple, and not very attractive.

The first example comment is clear about the nature and location of a problem, and it provides a strategy for addressing the problem. The second example comment gives a vague description of a problem, without any explanations or suggested solutions. When carefully hand-coded in a single course for a single assignment, research has found that features such as explanations and constructive suggestions in a comment can have benefits for comment recipients or comment providers (Deiglmayr, 2018; Wichmann et al., 2018).

In this study, to allow for analysis of comments from many assignments across many courses, we simple use comment length: the total number of words provided by a student across all reviews completed on assignment *J* (*Length Provided*$_J$) and the total number of words received across all reviews on the student's submission for assignment *J* (*Amount Received*$_J$). As with the number of comments received, the length of comments received was treated as missing when the student failed to submit a document for an assignment.

Helpfulness of comments provided and received. In *Peerceptiv*, as with several other popular online peer review systems (e.g., *CrowdGrader*, *MobiusSLIP, Peergrade. io*), feedback recipients are asked to judge the helpfulness of the feedback they have received. In *Peerceptiv*, this takes the form of a 1–5 Likert rating for each comment dimension. For the example comments given in the comment length measure description, the first comment received a 5 and the second received a 1 (and the optional comment to justify this low helpfulness rating: "*How is it too simple and not 'attractive'*?")

Only one helpfulness rating is given for the dimension in a given review, even if multiple comments were entered. Since we are aggregating to produce mean helpfulness across all reviews received by a student for assignment *J* (*Helpful Received*$_J$) or all reviews provided by the student for assignment *J* (*Helpful Provided*$_J$), this discrepancy is not problematic. Students are not shown the connection between specific ratings they received and the specific comments they received on their document, so the helpfulness ratings are primarily driven by the comment characteristics (e.g., was it understandable, was it actionable, did the author agree with the recommendation; Wu & Schunn, 2020a). The mean helpfulness ratings for a given assignment ignored missing helpfulness ratings unless none of the relevant ratings were completed, more commonly the case for helpfulness received since a student tended to complete all or none of the helpfulness ratings for a given assignment (mean %missing = 2%).

Assignment number. Since there could be temporal effects within a course that are confounded with independent and dependent variables (e.g., later assignments happen to produce more comments and show greater task gains), the relative assignment number within a course, *J*, is used as a covariate in the analyses.

3.2. Analysis

The general approach used is meta-regression (Harbord & Higgins, 2008; Thompson & Higgins, 2002) applied to cross-sectional, correlational, within-course data, in which multiple regression analyses are first conducted separately within each course, and then meta-regression is applied to the regression results to synthesize findings across courses. Each multiple regression within a course used a temporally-lagged model (Barnett et al., 2005), with relative task performance on assignment *J* being predicted by relative task performance on assignment *J–1* as a baseline control and by each of the review experience variables for assignment *J–1*. More specifically, in testing research questions 1, 2, and 3, we examined whether the amount, length, and helpfulness of feedback provided in the prior assignment and the amount, length, and helpfulness of feedback received in the prior assignment predict relative growth/decline in task score in the current assignment. Note that for several of the predictors, negative relationships are possible (e.g., demotivation from too much negative feedback) and thus two-tailed statistical thresholds are used.

The dependent variable, task score, is a continuous variable and generally was found to have a roughly normal distribution in each course. To further test the assumed linearity of relationships between predictors and the outcome variable, predictor variables were transformed in three equal-width categorical levels (within each course), and the relationships between the outcome and categorical variable level while controlling for other continuous variable correlates were estimated with separate regression models and plotted. Linear relationships were found in every case. Thus, linear models were selected for the multiple regressions.

The total number of data points in each multiple regression analysis for a given course is equal to the number of students in the course multiplied by the total number of assignments minus 1 (since there is no predictor of change for the first assignment). These within-course analysis *N*s varied between a low of 276 and a high of 1,646, with a mean *N* of 828 across the courses. Thus, all of the courses had sufficiently large *N*s to robustly evaluate the prediction strength of eight predictor values, and many of the courses had sufficient *N*s to produce effect size estimates with small error bars. Multiple collinearity issues were tested by examining Variance Inflation Factors, and no problematic values were found (i.e., all VIFs < 2.3).

To synthesize findings across course-specific applications of a given regression model, we conducted the meta-regression analyses in Stata version 15 using the random effects model with the "Metan" command. For every predictor in the multiple regression model, meta-regression produces a mean and 95% confidence interval for the effect size, along with *Z* and *p*-value that test the statistical significance against a null hypothesis of an effect size = 0. A t^2 represents the variance across courses in the observed effect sizes. This heterogeneity in effect sizes is formally tested with a χ^2 test and corresponding *p* value. I^2 provides an estimate of the proportion of variation across courses that is likely to be real (as opposed to random variation due to statistical impression). The combination of effect size confidence intervals and inferential tests of variation in effect sizes across courses serve as formal tests of research question 4. I^2 are interpreted as follows: 0%–40%: might not be important; 30%–60%: may represent moderate heterogeneity; 50%–90%: may represent substantial heterogeneity; 75%–100%: considerable heterogeneity (Thompson & Higgins, 2002).

The initial models included each of the reviewing experience measures as the main effects predictor, along with prior task score as the baseline and assignment number as the context. Follow-up analyses added several key interactions: interactions of amount and length with round (to test diminishing returns with more reviewing experience), interactions of amount and length with helpfulness (to test whether quantitative effects were larger when quality was taken into account), and interactions of length/amount provided with length/amount received.

Since the predictor of length provided produced the largest and most robust effect, two sets of follow-up regressions were conducted focusing on this predictor. First, additional regressions were conducted to test whether length provided as a predictor was suppressing the amount provided as a predictor by replacing total words across reviews with the mean length of each comment. If comment length is indeed the key predictor, the mean length provided should remain relatively strong, and the amount provided should remain relatively weak as a predictor.

Second, lag-two regressions were conducted to test the assumption that most of the learning from reviewing was occurring from the prior round. That is, the relative predictive strengths of provided length in the prior assignment and in two assignments ago were tested. If learning effects tend to be expressed immediately, then the lag-two predictor should show relatively weak predictiveness, and the lag-one predictors should show relatively strong predictiveness. These lag-two correlations also provide stronger additional controls for confounded third-variable factors related to students who tend to provide long comments. That is, perhaps students who provide long comments are generally more conscientious students who then work harder in the course to improve their performance, and the relationship between length provided to gains in task score merely reflects the relationship of both factors to conscientiousness. If that were true, then both lag-2 and lag-1 length should equally predictor growth in task score.

Finally, the inclusion of comment helpfulness in the regression models might have controlled for too much in estimating the relationship of comment length (provided or received) and task growth. Comment length is often an important part of perceptions of comment helpfulness: a very short comment simply does not have the content required to be helpful to the author. Thus, helpfulness might be considered a partial mediator of comment length's benefits for task growth. To address possible reductions in effect sizes, the models were re-run excluding comment helpfulness (received or provided) as predictors.

4. Results

Fig. 2 summarizes the key meta-regression findings across main effect and interaction multiple regression models. Table 3 presents the estimated effect size for each reviewing experience predictor of task growth from the main effect meta-regression (see Appendix Table A2 for regression estimates in each separate course). As a sanity check, prior task score was the strongest predictor of current task score in every course; thus, modeling student growth as change relative to the prior assignment generally made sense in these courses.

Among the review experience predictors, four had a significant overall effect size. These effect sizes are standardized betas. For example, an effect size of 0.147 means that for every one standard deviation increase in the predictor (*Length Provided*$_{J-1}$ in this case), there is a 0.147 standard deviation increase in the task score. These four overall effects were consistently obtained when the total provided length was replaced with the mean provided length (see Appendix Table A4 for course-specific coefficients) and when both lag-1 and lag-2 measures of provided length were included (Appendix Table A5). Further, the patterns for amount and length were robust when helpfulness predictors were removed (Appendix Table A6).

Table 4 presents the results of the interaction models (Appendix Table A3 presents the course-specific coefficients). Note that the key main effects were still significant overall when including interaction terms. Only two interactions were statistically significant: helpful provided x length received (a positive interaction) and the amount received x amount provided (a negative interaction). In the next sections, we interpret the patterns in these figures and tables in terms of each research question.

4.1. RQ1. Predicting task gains from comment amount vs. comment length

For both provided and received comments, comment length was found to be a stronger and a more consistently positive predictor of future task performance in comparison to the predictiveness of the amount of comments (see Fig. 2 and Table 3). These differences were substantial: the length was statistically significant overall for both provided and received; the amount was not statistically significant overall for either provided or received. Further, the robustness test involving replacing the total amount of words provided with the mean number of words as a predictor still found a larger overall effect of comment length relative provided to comment amount provided, although the two factors became more similar in effect size. Note, however, that the mean length predictor should not be considered a better measure of comment depth since it does not capture the total amount of learning through commenting provided by someone contributing many long comments rather than only a few long comments. Instead, this robustness test is

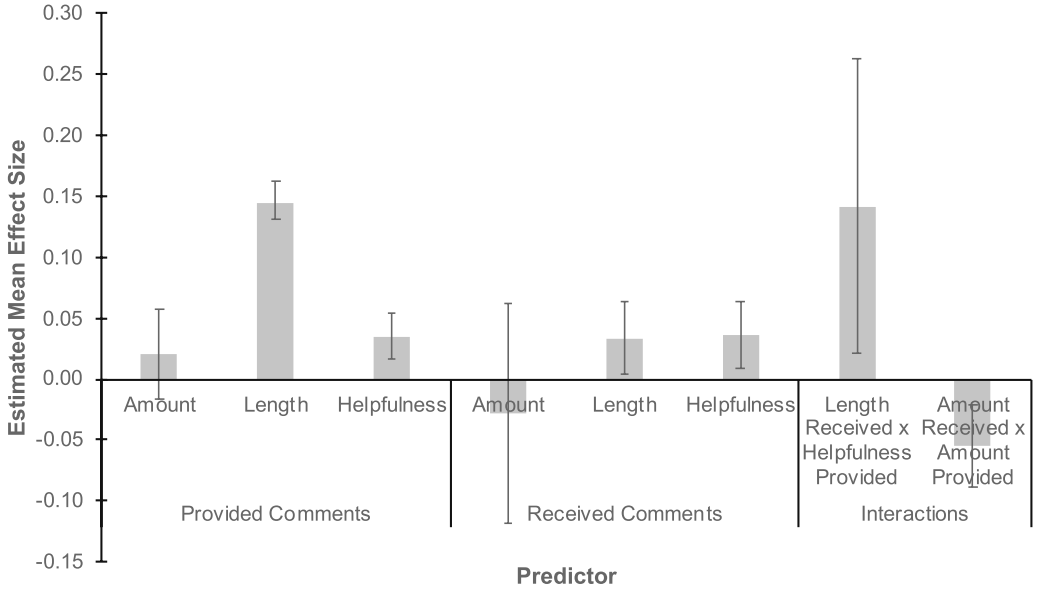

Fig. 2. Estimated mean effect size (with 95% confidence intervals) for each reviewing experience predictor of growth in future task performance.

Table 3
Meta-analysis results for the overall effect across courses of each core predictor of growth in task performance along with heterogeneity of effects across courses (statistically significant overall effects and effect heterogeneity are in bold).

	Overall Effect						*Heterogeneity of Effect*		
	ES	*ES 95% CI*		Tau^2	Z	p	χ^2	p	I^2
Baseline									
Task Score $_{J-1}$	0.200	0.167	0.254	0.0047	**9.25**	**<.0001**	**53.39**	**<.0001**	78%
Core predictors									
Amount Provided$_{J-1}$	0.020	−0.016	0.057	0.0043	1.08	.28	**1092**	**<.0001**	99%
Length Provided$_{J-1}$	0.147	0.131	0.163	0.0009	**17.7**	**<.0001**	**1.8e + 06**	**<.0001**	100%
Helpful Provided$_{J-1}$	0.035	0.016	0.055	0.0003	**3.63**	**<.0001**	15.66	.21	23%
Amount Received$_{J-1}$	−0.028	−0.119	0.062	0.0263	0.61	.54	**6256**	**<.0001**	100%
Length Received$_{J-1}$	0.033	0.004	0.063	0.0029	**2.20**	**.03**	**2.0e + 06**	**<.0001**	100%
Helpful Received$_{J-1}$	0.036	0.008	0.063	0.0003	**2.57**	**.01**	13.50	.33	11%
Assignment number	−0.014	−0.039	0.012	0.0013	1.02	.31	**33.65**	**.001**	64%

Table 4
Meta-analysis results for the overall effect across courses of each interaction predictor of growth in task performance along with heterogeneity of effects across courses. Main effect predictors are included in the models, but their values are not shown. Statistically significant overall effects and effect heterogeneity are in bold.

	Overall Effect						*Heterogeneity*		
	ES	*ES 95% CI*		Tau^2	*Z*	*p*	χ^2	*p*	I^2
Interactions with Assignment Number									
Amount Provided$_{J-1}$	0.011	−0.066	0.089	0.0203	0.28	.776	5780.55	**<.0001**	100%
Length Provided$_{J-1}$	0.011	−0.128	0.151	0.0653	0.16	.873	7083.59	**<.0001**	100%
Amount Received$_{J-1}$	−0.043	−0.109	0.023	0.0146	1.28	.199	3.1e+07	**<.0001**	100%
Length Received$_{J-1}$	0.091	0.001	0.161	0.0217	1.98	.048	2.0e+07	**<.0001**	100%
Interactions with *Helpful Received$_{J-1}$*									
Amount Provided$_{J-1}$	0.003	−0.200	0.205	0.1388	0.03	.980	15,076.8	**<.0001**	100%
Length Provided$_{J-1}$	−0.058	−0.257	0.140	0.1332	0.58	.563	1.0e+08	**<.0001**	100%
Amount Received$_{J-1}$	−0.137	−0.311	0.036	0.0998	1.55	.121	10,890.05	**<.0001**	100%
Length Received$_{J-1}$	−0.052	−0.185	0.002	0.0602	0.76	.449	2.1e+07	**<.0001**	100%
Interactions with *Helpful Provided$_{J-1}$*									
Amount Provided$_{J-1}$	−0.001	−0.122	0.12	0.0492	0.02	.988	11,184.2	**<.0001**	100%
Length Provided$_{J-1}$	−0.042	−0.245	0.168	0.1391	0.41	.683	3.0e+08	**<.0001**	100%
Amount Received$_{J-1}$	−0.016	−0.137	0.104	0.0478	0.27	.791	1160.28	**<.0001**	100%
Length Received$_{J-1}$	0.142	0.022	0.263	0.0492	**2.31**	**.021**	3.9e+07	**<.0001**	100%
Received × Provided Interactions									
Amount Received$_{J-1}$ x *Amount Provided$_{J-1}$*	−0.055	−0.089	−0.021	0.0035	**3.13**	**.002**	76,921.32	**<.0001**	100%
Length Provided$_{J-1}$ x *Length Received$_{J-1}$*	0.025	0	0.05	0.0021	1.92	.054	1.3e+12	**<.0001**	100%

more appropriately conceptualized as an especially strong lower-bound estimate of the relative contributions of depth vs. amount.

Note that the relative strengths of length vs. amount as predictors are likely under-estimated in these models because they included helpfulness as a predictor, and helpfulness is a likely mediator of comment length benefits. When helpfulness was removed as a predictor (for both provided and received comments), the mean effect size for provided length increased from 0.14 to 0.16, whereas the other effect size estimates remained unaffected (Appendix Table A6 presents course-specific coefficients).

4.2. RQ2. Predicting task gains from student perceptions of comment value

Student perceptions of comment helpfulness were statistically significant overall predictors of task learning, although with small effect sizes for both provided and received comments (see Fig. 2 and Table 3). Thus, student perceptions of comment value do appear to be meaningful predictors. It should be noted that the mean helpfulness ratings were generally high (typically around 4 on the 1–5 scale; see Appendix Table A1), which then necessarily produces relatively low variance in perceived helpfulness. It is also important to note that the relative predictive strength of provided comment helpfulness appears to increase substantially when taking into the length of comments received (see Fig. 2 and Table 4).

4.3. RQ3. Predicting task gains from received vs. provided comments

Quantitative aspects of both providing and receiving comments were significant independent predictors of growth in task performance, as one would have expected from the extant literature of the role of peer feedback in student learning. However, there is also strong support for the greater role of providing over receiving, at least with respect to the aspect of commenting that is more strongly associated with learning: comment length. That is, provided comment length, in particular, was a very strong predictor of growth in task performance, over four times as large an overall effect size compared with received comment length or any feature of received comments. The amount of comments was not a statistically significant predictor overall for either provided or received, whereas comment helpfulness had a similar and small but statistically significant effect size for both received and provided comments. These findings held across the various regression models that were tested.

It should be noted, however, that there was a statistically significant and large overall effect size for the interaction of length of *received* comments and helpfulness of *provided* comments (see Fig. 2 and Table 4). That is, it appears that long comments received are especially useful for length when they are received by students who provide more helpful comments to their peers. There is also a small negative interaction between the amount received and the amount provided. That is, it appears that the negative association of receiving more comments with future task performance is especially large for students who provide many comments.

4.4. RQ4: variation of predictive relationships across courses

The overall effect size estimates were useful for addressing the first three research questions, but they also mask considerable variation in each of the predictive relationships across the studied courses. As shown in Table 3, there was significant heterogeneity of effects (i.e., variation across courses that was not simple statistical imprecision in each course's effect estimate) for all but two of the key prior reviewing experience predictors (the two comment helpfulness predictors).

To visualize the variation in effects across courses, Fig. 3 presents the estimated effect size within each course for the four predictors that have statistically significant heterogeneity. While all of these predictors have statistically significant and large (in terms of I^2) heterogeneity of effects, there are two very different patterns. The variation in effect size for length provided is entirely quantitative. That is, in every single course, the effect size is substantial, positive, and precisely estimated; however, the specific effect sizes vary across courses in a 2:1 way (some as high as 0.2 and some as low as almost 0.1). By contrast, the variation in the effect sizes for the other three predictors is qualitative. That is, the effects are sometimes positive and substantial, but sometimes the effects are near zero/not statistically significant, and some are even negative. The variations in effects in this qualitative way are shown in Fig. 4, which shows what percentage of courses show a meaningful positive effect size, what percentage show a positive but very near zero effect size, and what percentage of courses show a negative effect size.

Another important point about course variation is that there is no simple pattern by course discipline. Even for the same level course in a given discipline at a given university (e.g., Biology 2–4 or Entomology 1–4), effect sizes can vary substantially. This suggests the variation in effect sizes depend upon how the assignments are structured, rather than being due to differences in the overall topics to be learned or the population of learners.

A third important pattern revealed in Fig. 3 involves the negative predictors. Particularly for the amount received and length received, there were four to five courses in which there were robust negative effect sizes: that is students who received more comments and longer comments tended to do worse in the next assignment. Given the small error bars on most courses' effect size estimates for these variables, this variation is very unlikely to be due to by-change variations in imperfectly measures relationships (as further supported by the formal heterogeneity tests). In the general discussion, we take up potential explanations of these negative relationships (e.g., demotivating feedback).

To further explore this issue of measurement imprecision in some of the course-specific estimates, we examine the by-course correlation (i.e., $N = 13$) among the different effect size estimates (e.g., did a course with a strong length provided effect size also tend to have a strong length received effect size; see Table 5). We also include two potential factors related to noise of estimates: course size (since few students should produce less noisy effect size estimates) and the number of dimensions in the peer reviewing rubric (since more dimensions should produce reliable task estimates). None of the correlations were especially large, and they even varied in their direction from positive to negative. Thus, there is no support for the idea that some courses had more precise estimates of learning and that the variation in effect sizes simply reflects that variation in the precision of measuring learning effects. Further, larger courses and courses with more reviewing dimensions did not generally produce larger effect size estimates.

5. Discussion

5.1. Conclusions

A large dataset of online peer feedback was leveraged to robustly test four major open research questions about the nature of peer feedback behaviors in task learning. In terms of overall estimated effect sizes across the courses, the meta-regression produced straightforward answers to the four research questions. First, task performance gains are much more strongly associated with comment depth rather than the number of comments. Indeed, for received comments, the amount of comments as a predictor even trended as a negative relationship to future task performance. From a cognitive process perspective, providing peer feedback involves a number of complex cognitive processes that should support learning (van Popta et al., 2017), but it now appears that students must deeply engage with these cognitive processes in order to reap the learning benefits.

Second, this study provided the first robust evidence that student perceptions of the value of provided comments tap important information regarding the learning potential from these comments. These predictive relationships held even when comment length was controlled. In other words, students seem to be able to make important judgments, at least at the university level, about whether comments are useful for learning even beyond flagging very short comments as non-productive. Thus, while the collection of these judgments in practice may often be about improving accountability for higher quality reviewing, from a research perspective, there is an important signal in this data related to learning. From a self-regulated learning perspective (Panadero et al., 2017; Winne, 2010), since these judgments appear to have some validity, it is possible that engaging students in this kind of evaluation of comment quality may further support student learning.

Third, the meta-regression strongly supported the much stronger role of providing comments over receiving comments as predictors of task performance growth. While both receiving peer feedback and providing peer feedback had statistically significant relationships to growth in task performance, the providing side appears to have a consistently greater potential for task performance growth. Although there was some emerging evidence from a few studies that tried to tease apart the providing and receiving sides of peer feedback for task learning (e.g., Lundstrom & Baker, 2009; Wu & Schunn, 2021), the current study shows this is indeed a general pattern with substantial differences in effect sizes.

Finally, by leveraging a dataset in which a fixed technology was applied across courses with a very specific shared approach to construct measurement, this study was able to careful document quantitative (and qualitative!) variation in effect sizes that was not simply measurement noise or comparison of apples to oranges at the construct measurement level. Almost every predictor showed large variation in effect sizes, and simple statistical noise explanations were ruled out. Further, some of the variations were found to be entirely quantitative (i.e., a significant positive relationship was found in every course, but the effect size varied across courses), whereas some of the variations were qualitative (i.e., several variables were sometimes significant positive predictors and sometimes significant negative predictors). Both forms of variation can be important both theoretically and practically, and they should be the focus of future research.

5.2. Limitations

The most obvious limitation to discuss involves causal inference from correlational data. The research questions were carefully posed to be about prediction rather than causation, but underlying theory and practice is especially interested in causal relationships. The temporally-lagged approach that also includes many covariates, along with the robustness of the outcomes across different statistical models, rules out reverse causality and many plausible confounding factors. Further, the causal status of received feedback in general, received peer feedback, and provided peer feedback is not currently in doubt, having been well established across a wide number of experimental studies that are well summarized in meta-analyses (Zheng et al., 2020; Huisman et al., 2019; Li et al., 2020; Double et al., 2019). Thus, it is very likely that the predictive relationships examined here are tapping causal relationships. However, it is not yet well established that quantitative variation in

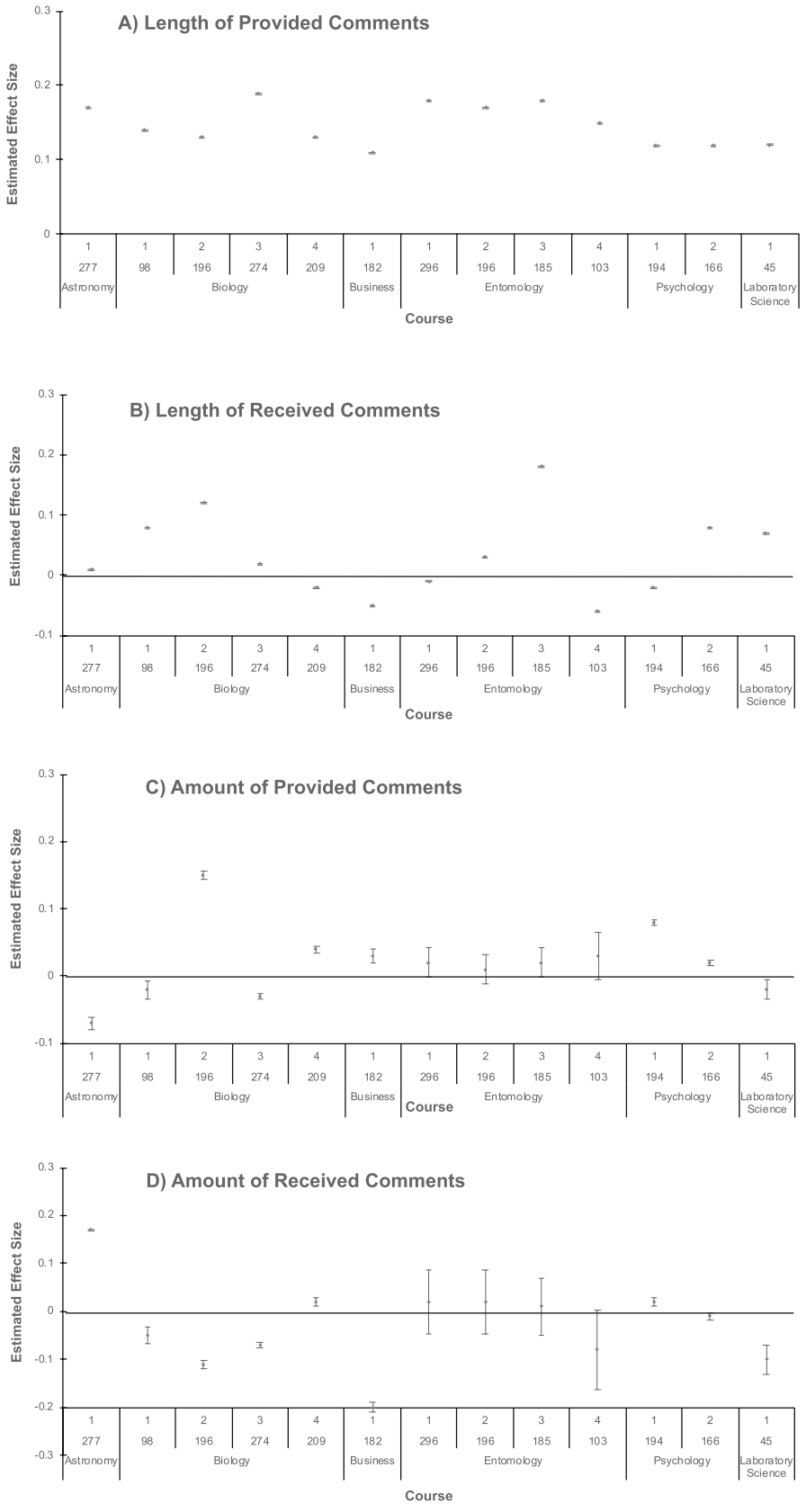

Fig. 3. Estimated effect size (with SE bars and course *N*s) on growth in task performance within each course of A) length of previously provided comments, B) length of previously received comments, C) amount of previously provided comments, and D) amount of previously received comments.

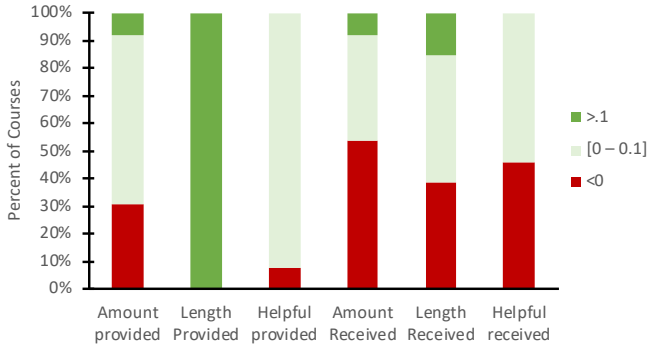

Fig. 4. Relative frequency of effect size categories across courses for each predictor of future task performance.

predictive strength closely corresponds to quantitative variation in causal strength. The patterns revealed by the current meta-regression results provide clear directions for particular effects that are worth testing experimentally. In addition, we note there are some assignment-level and learner-level variables that may have been confounded with our tested predictor variables and that future research should directly address: changing motivational levels across assignments in the course; changing assignment and rubric details across the course; accuracy of comments; and overlap (or disagreement) across reviews for one document.

The second limitation to acknowledge is a consequence of the research strategy of using one online peer feedback system to reduce spurious sources of variance. The specific supports and structures found in *Peerceptiv* will vary from those found in other online peer feedback systems (e.g., ranking vs. rating prompts, comments made directly on submitted documents vs. in textboxes, ways of allocating reviewers to documents, whether and how back-evaluations are implemented). Replicating the current findings across systems are needed to fully test the robustness of relationships observed here.

The third limitation involves educational contexts. The selected courses varied widely across disciplines and geographic regions in the US, with corresponding large variation in gender and race/ethnicity profiles. However, these were all university undergraduate courses, predominantly large enrollment courses, and all were in the US.

Finally, in applying a methodology to a large number of learners across many assignments, the measurement approach used simple counting strategies of easy-to-count variables and tested linear relationships. Although these approaches were grounded in prior work and revealed important patterns, it will be important for future work to complement the current gain-size of analysis with studies done with more fine-grained features of peer feedback (e.g., what kinds of features are included in longer comments). Future developments in Natural Language Processing technologies may allow for sufficiently robust classification of comment features to examine fine-grained commenting behaviors in large datasets (e.g., Leijen & Leontjeva, 2012; Misiejuk et al., 2020; Nguyen et al., 2017).

5.3. Theoretical implications

If the observed relationships are indeed causal, then there are a number of important implications for theories of learning from peer feedback. First, theories need to explain why providing (long) comments offers such large affordances for learning and why the receiving side is typically small and often not beneficial. It may be, for example, that the inherently passive nature of receiving feedback vs. the inherently constructive nature of providing feedback needs to be given a central role in theoretical accounts.

Second, theories need to explain the negative relationships, which likely will involve collecting additional data. Is there a motivational explanation involving students disengaging from the class when they receive a lot of criticism (Chen & Jang, 2010; Mega et al., 2014) or they provided more feedback than they received? Is there a competition for time devoted to a course between working on other aspects of the class that support task learning (e.g., reading or other assignment completion) and participation in peer feedback? Does attending to too many performance dimensions at once dilute the focus a learner requires to make rapid progress (Ericsson, 2004; Ericsson et al., 1993)?

Third, theories need to be expanded to account for variation in quantitative and qualitative variation in effect sizes. Theorizing in education is frequently restricted to simple qualitative predictions, given the complexity and breadth of factors that interact to produce learning outcomes. However, effect sizes have very serious implications for practice, and theories need to be improved to address the substantial qualitative and quantitative variation that was observed in the current study. Future studies that uncover the sources of this variation are needed. Potential explanations might include: 1) relative heterogeneity in student knowledge and skills which lead some learners to become demotivated by comparison to their peers' performance when there are large differences; 2) instructor supporting strategies for maintaining growth mindsets and overall motivation levels in the presence of detailed feedback for how to improve; 3) level of detailed support in evaluation rubrics for coherent, clear, and actionable peer feedback; and 4) the match of the assignment and evaluation rubric to typical student attitudes towards whether mastery of the underlying skills is worthwhile.

Research on peer feedback has moved over the last two decades from a small field that is narrowly focused on rating accuracy and feedback quality to a very large and interdisciplinary field that has produced many meta-analyses and synthetic reviews. The meta-analyses generally find significant heterogeneity of effects, which the current study has shown not to be attributable to measure variation across studies. The relative importance of traditional studies each done within just one course is therefore becoming much smaller. Instead, the focus of future peer feedback research must either pursue new variables or take up the task of describing and explaining variation in effects across contexts.

5.4. Practical implications

The current findings suggest instructors should try a variety of methods that encourage students to give long comments and discourse giving many small comments. This could be done through reviewing prompts that limit the number of comments that are made and include

Table 5
Correlation across (N = 13) courses among the estimated effects of the relationship of different reviewing experiences with growth in task performance (amount of comment provided, length of comments provided, amount of comments received, and length of comments received).

	Course size	# of dimensions	Task score	Amount provided	Length provided	Amount received
# of dimensions	−0.06					
Task score	−0.25	−0.35				
Amount provided	−0.17	−0.06	−0.27			
Length provided	0.36	−0.08	0.26	−0.42		
Amount received	0.27	0.08	0.22	−0.37	0.46	
Length received	−0.39	−0.03	0.22	0.10	0.14	0.05

detailed guidance on what might be included in a comment. Alternatively, this goal could be achieved through training provided to students on effective peer feedback practices (Nicol & Macfarlane-Dick, 2006; Nicol & Milligan, 2006). Finally, additional automated support tools could be adding into online peer feedback systems that monitor comment quality and prompt reviewers to expand low-quality comments (e.g., Ramachandran et al., 2017).

Author credit statement

Zheng Zong was responsible for data curation, conducting the data analyses, and writing the original draft. Christian Schunn was responsible for conceptualization of the research questions, design of the analytic models, design of the visualizations, reviewing/editing the drafts, and assisting in data curation. Yanqing Wang was responsible for project administration and reviewing/editing the drafts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2021.106924.

References

Applebee, A., & Langer, J. (2011). *The national study of writing instruction: Methods and procedures*. Albany, NY: Center on English Learning & Achievement. Retrieved December 27, 2011.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220.

Baturay, M. H. (2015). An overview of the world of MOOCs. *Procedia - Social and Behavioural Sciences, 174*(1), 427–433.

Bijami, M., Kashef, S. H., & Nejad, M. S. (2013). Peer feedback in learning English writing: Advantages and disadvantages. *Journal of Studies in Education, 3*(4), 91–97.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290.

Chen, K. C., & Jang, S. J. (2010). Motivation in online learning: Testing a model of self-determination theory. *Computers in Human Behavior, 26*(4), 741–752.

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

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

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.

Curran, F. C., Fisher, B. W., Viano, S., & Kupchik, A. (2019a). Why and when do school resource officers engage in school discipline? The role of context in shaping disciplinary involvement. *American Journal of Education, 126*(1), 33–63.

Curran, V., Gustafson, D. L., Simmons, K., Lannon, H., Wang, C., Garmsiri, M., & Wetsch, L. (2019b). Adult learners' perceptions of self-directed learning and digital technology usage in continuing professional education: An update for the digital age. *Journal of Adult and Continuing Education, 25*(1), 74–93.

Davey, K. R. (2011). Student peer assessment: Research findings from a case study in a master of chemical engineering coursework-program. *Education for Chemical Engineers, 6*(4), 122–131.

Deiglmayr, A. (2018). Instructional scaffolds for learning from formative peer assessment: Effects of core task, peer feedback, and dialogue. *European Journal of Psychology of Education, 33*(1), 185–198.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*, 481–509.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70–S81.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

van Gennip, N. A., Segers, M. S., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290.

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

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Harbord, R. M., & Higgins, J. P. (2008). Meta-regression in Stata. *STATA Journal, 8*(4), 493–519.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

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

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.

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

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.

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods, 10*(1), 83–98.

Leijen, D. A., & Leontjeva, A. (2012). Linguistic and review features of peer feedback and their effect on the implementation of changes in academic writing: A corpus based investigation. *Journal of Writing Research, 4*(2), 178–202.

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

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

Lu, J., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science, 40*(2), 257–275.

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

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168.

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal Of Educational Psychology, 106*(1), 121–131.

Min, H. (2016). Effect of teacher modeling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing, 31*, 43–57.

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2020). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, 106658.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. *Handbook of Research on Educational Communications and Technology, 3*, 125–144.

Nelson, G. L., & Murphy, J. M. (1992). An L2 writing group: Task and social dimensions. *Journal of Second Language Writing, 1*(3), 171–193.

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.

Nguyen, H., Xiong, W., & Litman, D. (2017). Iterative design and classroom evaluation of automated formative feedback for improving peer feedback localization. *International Journal of Artificial Intelligence in Education, 27*(3), 582–622.

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

Nicol, D., & Milligan, C. (2006). Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice. In C. Bryan, & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 64–77). Taylor and Francis Group Ltd.

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, 40*(8), 1253–1278.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98.

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.

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

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.

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.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534–574.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565.

Schunn, C. D., Godley, A. J., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent and Adult Literacy, 60*(1), 13–23. https://doi.org/10.1002/jaal.525

Scott, T. M., Gage, N., Hirn, R., & Han, H. (2019). Teacher and student race as a predictor for negative feedback during instruction. *School Psychologist, 34*(1), 22–31.

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.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science, 29*(4), 549–571.

Sluijsmans, D. M., Moerkerke, G., Van Merrienboer, J. J., & Dochy, F. J. (2001). Peer assessment in problem based learning. *Studies In Educational Evaluation, 27*(2), 153–173.

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

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*(11), 1559–1573.

Topping, K. (1998). Peer assessment between students in college and universities. *Review of Educational Research, 68*(3), 249–279.

Traga Philippakos, Z. A., MacArthur, C. A., & Munsell, S. (2018). College student writers' use and modification of planning and evaluation strategies after a semester of instruction. *Journal of Adolescent & Adult Literacy, 62*(3), 301–310.

Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161–1174.

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

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

Wichmann, A., Funk, A., & Rummel, N. (2018). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *European Journal of Psychology of Education, 33*(1), 165–184.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*(4), 267–276.

Wooley, R. S., Was, C. A., Schunn, C. D., & Dalton, D. W. (2008, July). The effects of feedback elaboration on the giver of feedback. In *Paper presented at the 30th annual meeting of the cognitive science society, Washington, DC*.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699.

Wu, Y., & Schunn, C. D. (2020a). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology, 60*, 101826.

Wu, Y., & Schunn, C. D. (2020b). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemporary Educational Psychology, 62*, 101897.

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

Zhang, F., Schunn, C. D., & Baikadi, A. (2017). Charting the routes to revision: An interplay of writing goals, peer comments, and self-reflections from peer reviews. *Instructional Science, 45*(5), 679–707.

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.

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.