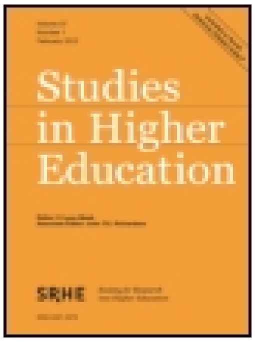

Studies in Higher Education

ISSN: 0307-5079 (Print) 1470-174X (Online) Journal homepage: http://www.tandfonline.com/loi/cshe20

# Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback

Melissa M. Patchan, Christian D. Schunn & Russell J. Clark

**To cite this article:** Melissa M. Patchan, Christian D. Schunn & Russell J. Clark (2017): Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback, Studies in Higher Education, DOI: 10.1080/03075079.2017.1320374

**To link to this article:** http://dx.doi.org/10.1080/03075079.2017.1320374

Published online: 02 May 2017.

STUDIES IN HIGHER EDUCATION, 2017
https://doi.org/10.1080/03075079.2017.1320374

Routledge
Taylor & Francis Group

# Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback

Melissa M. Patchan[a], Christian D. Schunn[b] and Russell J. Clark[c]

[a]Learning Sciences & Human Development, West Virginia University, Morgantown, WV, USA; [b]Learning Research & Development Center, University of Pittsburgh, Pittsburgh, PA, USA; [c]Physics & Astronomy, University of Pittsburgh, Pittsburgh, PA, USA

**ABSTRACT**

We examined the influence of accountability on the consistency of peer ratings and quality of peer feedback by comparing three conditions: only rating accountability, only feedback accountability, or both rating and feedback accountability. From a large undergraduate course, 287 students' peer ratings and peer feedback were coded for rating consistency, comment helpfulness, amount of feedback, and feedback features. Because only 30% of the students accurately perceived their assigned condition, data were analyzed according to the perceived condition. Students who believed their reviewing grade would be influenced by the helpfulness of their feedback not only provided more feedback, but also more criticism, solutions, and localized comments. These students also provided more consistent ratings than those who thought their reviewing grade would be influenced by the consistency of their ratings. These findings indicate that constructing helpful comments could have a broad influence on peer assessment and consistent ratings are grounded in commenting.

**KEYWORDS**

Peer assessment; peer review; peer ratings; peer feedback; accountability

Peer assessment (often also called peer review) is the quantitative evaluation and qualitative feedback of a learner's performance by another learner. It is typically implemented in classrooms with the intention of developing the knowledge or skill of all learners involved. This form of peer assessment combines summative assessment (i.e. peers evaluate an individual's work in order to assign a grade) and formative assessment (i.e. peers provide constructive feedback that could help an individual improve his or her work) rather than focusing on assessing contributions to a group assignment that is graded by an instructor.

Peer assessment is a rich task that addresses multiple goals. First, peer assessment facilitates the inclusion of writing assignments in large classes where an increase in the teacher's workload would make such assignments not feasible. Therefore, to ensure that the resulting grades produced by peers are fair, the reliability and validity of peer ratings must meet standards. However, the constructs of reliability and validity are complex, especially in peer assessment. For example, is one considering the reliability and validity of individual reviews or the collective reviews; is one considering the reliability and validity of finding problems, assessing their importance, or suggesting repairs? In addition, the relationship between reliability and validity cannot be ignored – that is, it is more difficult to be valid when the reliability is low.

Second, and perhaps more importantly, peer assessment provides unique learning opportunities for students. Multi-peer assessment can provide more total feedback than a single over-taxed

**CONTACT** Melissa M. Patchan ✉ melissa.patchan@mail.wvu.edu Learning Sciences & Human Development, West Virginia University, 355 Oakland Street, Allen Hall 504M, Morgantown, WV 26506, USA

instructor (Cho, Schunn, and Charney 2006; Patchan, Charney, and Schunn 2009; Patchan, Schunn, and Clark 2011). This feedback is critical to student learning – both from constructing helpful feedback (Topping et al. 2013) as well as by receiving helpful feedback (Shute 2008). However, how much students learn from these opportunities will likely depend on the quality of the feedback. One indicator that students are detecting important problems versus trivial problems is the reliability and validity of the related ratings.

The use of peer assessment is supported by more than four decades of research (Bruffee 1980; Elbow 1973; Moffett 1968). This research has demonstrated that students are capable of providing reliable and valid ratings (Cho, Schunn, and Wilson 2006; Falchikov and Goldfinch 2000). Moreover, students can provide feedback that is just as helpful as an instructor's feedback in helping their peers improve their drafts (Topping 2005), and sometimes they can provide feedback that is more helpful (Cho and MacArthur 2011; Cho and Schunn 2007; Hartberg et al. 2008). However, as with most pedagogy, these effects are not consistent, which warrants a deeper investigation into the features of peer assessment.

Given the relatively large number of online peer assessment systems currently available (e.g. a sample of 40 systems were recently identified in the literature and on the web; Babik et al. 2016), it is not surprising that there is an equally abundant and diverse set of methods that focus on improving review quality (see Table 1). These methods can be divided into two categories: (1) approaches that focus on the reliability and validity of ratings and (2) approaches that focus on the quality and helpfulness of comments. Because this area is a new research topic, we begin with a broad review of these approaches and discuss their advantages and disadvantages. We then present a study that contrasts one rating-focused approach and one comment-focused approach from the larger set of methods found across all of these systems.

## *Approaches that focus on the reliability and validity of peer ratings*

### *Reliability weights*

Peer ratings are often used as summative assessment, where these ratings are used to calculate a grade for the assignment. Given the importance of the peer ratings in these circumstances, both students and instructors have expressed concern that some students may not be motivated or knowledgeable enough to provide ratings that are consistent with their peers and the instructor (Kaufman and Schunn 2011). Several online peer assessment systems (e.g. Aropä, Calibrated Peer Review (CPR), Expertiza, SWoRD – also known as Peerceptiv) directly address these concerns by using reviewer weights. For example, Aropä uses an algorithm to indicate how well the ratings a reviewer provided matches the ratings provided by other peers who evaluated the same work (Hamer, Ma, and Kwong 2005). Song, Hu, and Gehringer (2015) demonstrated that these existing algorithms produce weighted grades that are more accurate than the naïve average of peer ratings. Thus, using the reviewer weight to calculate grades will likely decrease the influence of poorly motivated or confused students.

### *Reliability grades*

Although some online peer assessment systems only use the reviewer weights to weight the scores assigned to the authors, some researchers have recommend using the reviewer weight to reward reviewers who provided more accurate ratings (Hamer, Ma, and Kwong 2005). However, Hamer and colleagues observed a wide range of reviewer weights, which made it difficult for them to provide more specific suggestions to instructors about how to interpret the weights. By contrast, the SWoRD system assigns a reviewing grade that is partly based on the reliability of the peer ratings (Cho and Schunn 2007; Schunn 2016). This portion of the grade is calculated by comparing all the ratings for a particular student to the mean ratings provided by the other peers who reviewed the same documents. For ratings that are more similar, a higher reviewing accuracy grade is assigned, and for ratings that are less similar, a lower reviewing accuracy grade is assigned. Schunn (2016) posits that this reviewing accuracy grade will force student to take the rating task seriously.

**Table 1.** Advantages and disadvantages of existing approaches to improve review quality.

| Approach | Examples of technology-supported systems | Advantages | Disadvantages |
|---|---|---|---|
| Ratings | | | |
| Reviewer weight/ reputation systems/ accuracy grades | Aropä, CPR, Expertiza, SWoRD | • Decreases the influence of poorly motivated or confused students<br>• Increases students' motivation to rate accurately | • Does not improve students' ability to appropriately apply reviewing criteria |
| Calibration/training | CPR, Expertiza, SWoRD | • Increases students' confidence in rating ability<br>• Increases reliability and validity of ratings beyond the initial assignment | • Increased workload for students<br>• Students may negatively perceive a task for which no one receives the benefit of their feedback |
| Comments | | | |
| Minimum word count | PeerMark | • Increases the length of comments | • Some students may cheat the system and produce longer comments that are just wordy |
| Non-anonymous reviewing | CritViz | • Increases students' motivation to provide better reviews because their reputation is on the line | • Students are uncomfortable providing peers feedback if their identify is known and are less likely to provide honest feedback |
| Teacher overview | Eli Review, My Reviewers, PeerMark | • Increases students' motivation to provide helpful feedback | • Increased workload for instructors |
| Back-review, double-loop feedback, metareviewing | CrowdGrader, MobiusSLIP, Peer Grader, PECASSE, SWoRD, Eli Review | • Increases students' motivation to provide helpful feedback<br>• Provides feedback to help students understand why authors found their comments helpful or not | • Increased workload for students<br>• Students may make biased judgments about feedback helpfulness<br>• Students may not read their back-review comments |
| Automated metareview/feedback | Expertiza, SWoRD | • Provides feedback to help students understand what features of comments are theoretically useful and whether their comments include these features or not | • The detection of some important features/content may not be easily automated |
| Training | SWoRD | • Increases the helpfulness of feedback beyond initial assignment | • Increased workload for students |

### *Training on ratings*

Despite the advantages of using reviewer weights or reviewing accuracy grades, these approaches do not directly focus on students' ability to appropriately apply the reviewing criteria, especially if many students, as novices, have misconceptions related to rating dimensions. To address the validity of ratings, several online peer assessment systems have incorporated a calibration or training component (e.g. CPR, Expertiza, SWoRD).

The details of the calibration process vary across systems. In CPR, each student must successfully complete a calibration task before rating the assigned peer documents (Balfour 2013; Russell 2004). This calibration task involves rating three essays that the instructor chooses strategically. The students receive a reviewer weight that is based on how well their ratings match the instructor's ratings for the same essay. They also receive written feedback about whether their rating was correct or why it was incorrect. If a student has not met a specified benchmark, then he or she must reattempt the calibration task until the performance is satisfactory. In SWoRD, students complete a training assignment that uses the large number of peer ratings each document receives to calculate a 'correct' or expert evaluation rather than a single expert rating that is likely to be noisy. SWoRD also only requires a fixed single round of training rather than requiring multiple iterations until a student has demonstrated satisfactory performance. In both CPR and SWoRD, the calibration task is completed before the reviewing assignment (i.e. stand-alone calibration). Other online peer assessment systems have combined the calibration task with the reviewing assignment (i.e. mixed calibration). For example, in Coursera, one of the five submissions assigned for review were also graded by an instructor (Piech et al. 2013). In Expertiza, students reviewed two sample products that were intentionally created to emphasize common mistakes along with one peers' product (Song et al. 2016).

These calibration and training tasks could benefit students in several ways. First, after completing the task, students may feel more confident in their rating ability. This confidence could lead to more accurate ratings as well as more favorable perceptions of peer assessment in general. Although the primary purpose of these calibration and training tasks is usually to calculate an accurate reviewer weight, the feedback students receive would likely improve their rating ability and allow them to rate more accurately on future assignments even when the calibration or training task is no longer required. One disadvantage to the calibration and training tasks is the increased workload for the students. As in all cases where the workload increases, the inclusion of such tasks may inhibit instructors to include peer assessment tasks because they might feel their students would be overburdened with tasks that are not directly relevant to the course objectives. Additionally, students may feel annoyed by having to evaluate documents for which no one receives the benefit of their feedback.

## *Approaches that focus on the quality and helpfulness of peer feedback*

Peer feedback is often used as formative assessment, and students are expected to revise (and improve) their draft by responding to their peers' feedback. Despite the mounting evidence that peers can be just as helpful and sometime more helpful than instructors (Cho and MacArthur 2011; Cho and Schunn 2007; Hartberg et al. 2008; Topping 2005), students are often skeptical of the usefulness of feedback their peers produce because they are concerned that not all peers are capable of helping them (Kaufman and Schunn 2011). To increase the helpfulness of peer feedback, several approaches have been used.

### *Minimum comment length*

First, PeerMark (the peer assessment tool offered by turnitin.com) allows teachers to set a minimum word count, which prevents students from submitting feedback that is too short. The assumption behind this approach is that longer comments will contain more useful content, and thus be more helpful than shorter comments. Although this approach increases the length of the feedback, some students may game the system and produce longer comments that are just wordy or redundant rather than including additional substantive details that help the writer understand the problem or how to revise the text.

### *Public reputation*

Second, CritViz motivates students to produce helpful comments by making their comments public (Tinapple, Olson, and Sadauskas 2013). Whereas many systems make the author-reviewer assignment double-blind (i.e. reviewers do not know who the authors are that they are reviewing and authors do not know who their reviewers are), CritViz makes all students' work and feedback public after the review period is over. Although this method may increase students' motivation to provide better reviews because their reputation is on the line, some students may be uncomfortable providing peers feedback if their identity is known, and they may be less likely to provide honest feedback.

### *Teacher monitoring*

Third, several systems allow teachers to monitor the peer feedback and make adjustments to grades as needed (Eli Review, My Reviewers, PeerMark). For example, in Eli Review, instructors can respond to peer feedback in three ways (Grabill, Hart-Davidson, and McLeod 2012). They can 'endorse' a comment by either responding to the writer by selecting the message 'keep this in mind while revising' or responding to the reviewer by selecting the message 'keep writing high-quality feedback like this.' They also have an opportunity to add their own comments to the peer feedback. Knowing that their teacher will also be reading and evaluating the comments, students will likely be more motivated to write helpful feedback. However, this approach would also increase the workload for instructors, which could make instructors reluctant to include peer assessment assignments.

### *Author ratings of review helpfulness*

Fourth, several systems have incorporated back-reviews (also called double-loop feedback or metareviewing) – that is, authors evaluate the feedback that they receive (e.g. CrowdGrader, MobiusSLIP, Peer Grader, PECASSE, SWoRD, Eli Review). This evaluation may involve rating the quality and helpfulness of the comments or whether the author agreed with the feedback. For example, CrowdGrader (de Alfaro and Shavlovsky 2014) assigns a 'crowd-grade' that takes into account three separate grades that reflect the reliability and validity of the ratings as well as the helpfulness of the feedback. To calculate the helpfulness grade, students are first asked to rate and comment on the helpfulness of each review they receive. Then, an algorithm is utilized, which excludes the lowest rating and applies a greater weight for negative feedback than positive feedback. Students are assigned a lower helpfulness grade when more of their reviews were considered unhelpful than helpful.

Such back-evaluations can directly influence comment quality. For example, knowing that a grade will be assigned based on the helpfulness of their feedback, students are likely to be more motivated to write helpful feedback. Further, focusing reviewer attention on helpfulness may improve comment quality separate from the motivation effects. However, the back-review task increases the workload for the student, which could make peer assessment less appealing for instructors. Moreover, students' judgments about helpfulness may be biased. For example, if students feel like they are receiving a negative review, they might retaliate by marking the feedback as not helpful. Students may also not appreciate the helpfulness of a comment that points out a problem that will involve a lot of revision to fix.

However, the back-review task also has potential long-term benefits. In addition to rating the helpfulness of the feedback, students are also expected to provide feedback explaining why comments were helpful or not. This qualitative feedback about comment quality could help students write more helpful feedback in the future. Unfortunately, not all students read their back-review comments since they are typically not expected to revise their feedback.

### *Automated evaluation of comment quality*

Fifth, some systems have automated this back-review process. For example, in Expertiza, an automated metareview feature has been integrated, which automatically calculates the number of unique comments provided and whether a reviewer's comments (1) are relevant to a specific submission, (2) offer praise, describe a problem, or suggest a solution, (3) cover all the 'important topics,' (4) are positive or negative in tone, and (5) included plagiarism (Ramachandran 2013). The SWoRD system has a feature that automatically detects whether the comments students submit include solutions or localization (i.e. a description of where the problem occurred) and prompts students to revise their feedback if a pre-specified threshold has not been met (Nguyen and Litman 2014; Nguyen, Xiong, and Litman 2014; Xiong, Litman, and Schunn 2012). Similar to the back-review approach, this automated feedback likely increases the students' motivation to provide helpful feedback and provides feedback to help students become better reviewers, but it does so without increasing the workload for the students and the potential bias in ratings. Despite these benefits, the detection of some important features or content of feedback may not be easily automated.

### *Training on comment helpfulness*

Finally, SWoRD's training assignment is aimed at not only improving peer ratings, but also peer feedback. The training assignment uses a variation of back-reviews to increase students' understanding about how helpful a comment would have been if they were the author. This training assignment will likely benefit students for the long-term – that is, students will improve their ability to construct helpful feedback for future assignments. However, this training task will also increase the students' workload.

## *Current study and hypotheses*

As outlined in the previous sections, there are many approaches to improving the quality of peer assessment, with differences across systems in what aspects of peer assessment are evaluated, and whether the approach uses accountability, training, or simply feedback to motivate reviewers. Interestingly, only some of these methods make students accountable for their own work. Accountability for a given performance dimension occurs when a student is held responsible for the quality of completed work on that performance dimension, and may be especially important for insuring high-quality participation when students are given tasks that involve significant work like peer assessment. These explicit accountability approaches would include (1) grades for rating consistency based on reviewer weights, (2) reputation for comment quality by public reviews, (3) grades for comment quality based on teacher overview, and (4) grades for comment helpfulness by peer rating. In a systematic review of tools that support peer assessment, Luxton-Reilly (2009) observed that less than half of the web-based systems included some form of explicit accountability. Equally disappointing is the lack of empirical research examining the effects of these approaches on the quality of peer assessment. Therefore, the goal of the current study is to examine the effects of accountability on the consistency of peer ratings and quality of peer feedback. Two particular forms of accountability are chosen from the two main aspects of peer assessment (i.e. ratings and comments).

In the current study, students completed the peer assessment tasks using a web-based peer assessment environment, SWoRD (Scaffolded Writing and Rewriting in the Discipline) (Cho and Schunn 2007; Schunn 2016). This environment includes accountability for both the peer ratings and peer feedback – that is, the grade students receive for completing the reviewing tasks comprises two parts: rating consistency and comment helpfulness (see the Measures section for specific details). We compared three accountability configurations (i.e. only rating accountability, only feedback accountability, or both rating and feedback accountability). This method was based upon the maxim 'what you test is what you get.' As Schoenfeld (2007) pointed out, students use tests to determine what they need to know and in doing so, the test influences students' attention and what is learned. Similarly, which reviewing tasks are graded will likely influence where students direct their effort. This assumption led to our first hypothesis.

*Hypothesis 1:* According to the *direct accountability hypothesis*, the reviewing grades will directly affect the quality of peer assessment – that is, the rating consistency grades are expected to improve the consistency of peer ratings but not comment quality, while the helpfulness grades are expected to improve comment quality but not consistency of peer ratings.

However, rating the quality of one's work and providing constructive feedback likely involve different types of processing. For rating tasks, reviewers only need to detect problems, and do so as a severity or frequency judgment at an aggregate level to produce a rating. The processing necessary to complete this task is more consistent with the fast and intuitive type of processing described in dual-process theories of reasoning (Evans 2011). By contrast, to construct helpful comments, reviewers need to detect, diagnose, and solve problems and do so at the individual problem level. This task is more consistent with the second type of reflective thinking found in dual-process theories of reasoning. Indeed, such a slower and more deliberative process is likely to not only lead to more helpful comments, but after reflecting on their peers' errors, students are also likely to make more reliable judgments carefully grounded in evidence. Such a more deliberative process from commenting may explain why students asked to provide comments also improve their own performance more than those who only rated the quality (Hamer, Ma, and Kwong 2005; Wooley et al. 2008). Thus, it is possible that there is a secondary or additive effect of accountability – that is, by better understanding, the problems in a peer's text through this deeper processing, the reviewer might provide more consistent ratings as well. This assumption led to our second hypothesis.

*Hypothesis 2:* According to the *depth-of-processing hypothesis*, when held accountable via reviewing grades, the commenting task is expected to involve deeper processing that will also improve the consistency of the peer ratings – that is, the helpfulness grades are expected to improve comment quality, which in turn will improve the consistency of peer ratings.

There are multiple possible relationships between the two hypotheses. At a basic level, some form of direct accountability for commenting is built into Hypothesis 2: commenting grades must influence the quality of peer feedback to have a depth-of-processing effect. Further, it may be that both hypotheses are correct, in which case commenting changes only when commenting is graded, but review consistency improves through either rating consistency grades or comment helpfulness grades with the best rating consistency obtained when both elements are graded. Alternatively, it maybe that only Hypothesis 2 effects are seen: comment quality and rating consistency both improve when comment helpfulness is graded but there are no effects of rating consistency grades because deeper processing is required to improve consistency.

## Method

### *Course context and participants*

The participants in this study included undergraduate students enrolled in an Introduction to Laboratory Physics course at a top-tier mid-sized public research university in the US. This course aimed to teach students about how the experimental process works by engaging them in obtaining, analyzing, and presenting their own experimental results. The course was structured in two parts: a 50-minute lecture in which students were introduced to the physical principles, and a lab session in which students collected and analyzed the data. Students were enrolled in one of three possible lectures that were all taught by the same instructor and 1 of the 15 possible lab sessions that were taught by 10 graduate teaching assistants (TAs). In addition to the informal lab reports, quizzes, and final exam, students were required to write one formal lab report on one of the eight labs that they completed prior to the due date for the first draft. The formal lab report was structured like a journal article and included an abstract and sections that describe the introduction and theory, experimental setup, data analysis, conclusion, and references. This lab report and its peer review serves as the focal object of our experimentation and analysis.

Of the 317 students enrolled in the course, 13 students opted out of allowing their data to be included in this research study. Because the accountability manipulation involved grading procedures different from the default procedures used in SWoRD, data from 17 students who previously used SWoRD were also excluded from the analyses. These students might have already learned certain reviewing styles. Therefore, data from 287 students were included for analysis. This sample (52% female; $M_{\text{age}} = 21.2$ years; $\text{SD}_{\text{age}} = 2.8$) represented students at all undergraduate levels (6% freshman, 26% sophomore, 52% junior, 15% senior, 2% post-baccalaureate) and a variety of majors but with a predominance of natural science majors (82% natural science, 5% social science, 5% multiple disciplines, 3% humanities, 3% undeclared, 1% engineering; 1% business). A variety of ethnicities were also represented (66% Caucasian, 22% Asian, 4% African American, 4% Hispanic, 4% other).

### *Design and procedures*

Data from the formal lab reports and their peer review were collected and analyzed. This overall task was intended to mirror an authentic dissemination process. After completing their first draft, authors uploaded their papers to the SWoRD system (Schunn 2016). After the first draft deadline, four peers' papers were randomly assigned to each reviewer. Reviewers had two weeks to complete the reviews. This task was scaffolded with a detailed rubric that included general reviewing suggestions (e.g. be nice, be constructive, be specific) as well as guidelines for the specific reviewing dimensions. Reviewers were expected to rate the quality of the draft on 10 dimensions using a seven-point scale. For each rating, they were given descriptive anchors to help with determining which rating was most appropriate. Reviewers were also expected to provide constructive comments on six dimensions (although ratings and comments were generally paired, some comment dimensions had two separate rating dimensions corresponding to sub-aspects of the larger comment dimension). For each dimension, the reviewers were prompted with several questions that directed their attention to relevant aspects of the report. The reviews were released to the authors after the reviewing deadline, and the authors had two weeks to revise their draft based on the comments provided by their peers. As they submitted their final draft, authors rated the helpfulness of the peer feedback using a seven-point scale. The TAs graded the final drafts with the same rating scale used by peers to evaluate the first draft. These TA ratings were not used in the current study.

Each recitation section was randomly assigned to one of three possible conditions that manipulated for which aspects of the process students were accountable: both, ratings only, or comments only. Details about the writing and peer review tasks were given in the recitation session, posted as an announcement on the course's learning management system, and emailed to the students. The instructor repeatedly emphasized that the goal of the reviewing task was to help the authors write a better second draft. As determined by the instructor of the course, the formal lab report accounted for 10% of the students' final grade, of which 3% depended on the quality of the reviews they provided in order to hold them accountable for their reviews. To foreshadow the results, the condition labels reflect the condition to which students were *assigned* to contrast with the conditions students sometimes *perceived* themselves to be in. In the *assigned both* condition, a student's reviewing grade was based on the consistency of their ratings (i.e. the degree to which the reviewer's ratings were consistent with the ratings provided by the other peers who also rated the same paper) and the helpfulness of their comments (i.e. how helpful their comments were perceived by the authors). In the *assigned ratings only* condition, a student's reviewing grade was based solely on the consistency of their ratings. In the *assigned comments only* condition, a student's reviewing grade was based solely on the helpfulness of their comments. In all conditions, students were required to provide ratings and comments so that any differences between conditions could be attributed to accountability rather than the completed task.

To check whether students were aware of and remembered the manipulation condition to which they were assigned, they were asked to identify which factors influenced their reviewing grade via a survey given after completing the reviewing task. Participants responded 'yes,' 'maybe,' or 'no' to two items that corresponded to the rating and commenting forms accountability (i.e. whether my ratings are consistent with the ratings provided by the other peers who also rated the same paper; how helpful my comments are in helping my peers' write their second draft). Because 70% of students misperceived the form of accountability that was applicable to their assigned condition, students were also grouped into perceived conditions for analyses based on their responses to the two corresponding items – students who indicated 'yes' on both items were assigned to the *perceived both* condition, students who indicated 'yes' to only the item about the ratings were assigned to the *perceived ratings only* condition, and students who indicated 'yes' to only the item about the comments were assigned to the *perceived comments only* condition.

## Measures

We examined one measure of rating quality and multiple measures of comment quality to test the effects of accountability on the peer review process.

### Rating consistency

Each reviewer's *rating consistency* was determined using the SWoRD-generated reliability coefficient, which is based on the relative consistency of the reviewer's rating to the mean reviewing rating (excluding the reviewer's own ratings) across the same dimensions and documents (i.e. the Pearson correlation between reviewer and peer means across the $n = 40$ ratings – 4 papers × 10 rating dimensions).

### Comment helpfulness

The perceived helpfulness of comments was determined using the back-review ratings on a 1 (very unhelpful) to 5 (very helpful) scale. These ratings represent the author's perception of helpfulness for the comments they received. *Comment helpfulness* for each reviewer was computed as a mean of the received back-review ratings across all 24 comments they provided (4 papers × 6 comment dimensions/paper). Since helpfulness is subjective and potentially influenced by positivity of the feedback, additional measures of comment quality were also used.

### Amount of feedback

As a second kind of additional measure of comment quality, the amount of feedback was examined using three different measures. First, the *volume of feedback* was computed by summing the number of words across comments provided. This measure represents the integration of multiple dimensions (i.e. length per comment × number of comments). Second, the overall *number of comments* was computed by summing the number of comments provided across dimensions; within each comment dimension, reviewers could provide between one and five distinct comments. Third, the *number of long comments* was computed by counting the number of provided comments that consisted of 50 words or more (the threshold of 50 words was based on a frequency histogram, which revealed that a majority of the comments contained fewer than 50 words – that is, comments that were 50 words or more were especially long). Although it is assumed that longer comments will contain more helpful information, it is possible that these longer comments are just more verbose. Therefore, we also examine the content of the feedback in the next section.

### Feedback features

As a third kind of measure of comment quality, three useful feedback features of the thousands of review comments were automatically coded using a classification model derived from data mining and Natural Language Processing techniques (Xiong, Litman, and Schunn 2012). The classifier automatically detected whether the comment included criticism, a solution, and localization. The *number of criticism comments* (i.e. comments that described a problem or offered a solution), the *number of solutions* (i.e. comments that suggested a way to improve the paper), and the *number of localized comments* (i.e. comments that describe the specific location of the problem or where to apply a solution) was computed by counting the number of each feedback feature. By comparing the automatically detected features to hand-coded features, prior work has demonstrated that the classifier is fairly reliable (e.g. for criticism, Kappa = .59; for solution, Kappa = 61; for localization, Kappa = .62; Nguyen and Litman 2014; Nguyen, Xiong, and Litman 2014).

## Results and discussion

### *Manipulation check*

Of the 287 participants, 244 students completed the survey questions used to determine their perceived condition. In general, the manipulation was not perceived as intended – only 30% of the students' perceived condition exactly matched their assigned condition. The most common belief was that the reviewing grade was based on both rating consistency and comment helpfulness (40%). The next common belief was that the reviewing grade was based on comment helpfulness only (29%). Finally, only 11% of the students believed that their reviewing grade was based on rating consistency only. Interestingly, 49 (20%) students indicated that neither rating consistency nor comment helpfulness influenced the reviewing grade. The perceived conditions did not significantly differ in demographics. Moreover, there were no differences in their reported SAT scores, their reported freshman composition grade, or their prior experience with peer review.

Because the goal of this paper was to examine the effects of explicit accountability on the quality of peer assessment, the main analyses focus on the perceived conditions, excluding the 49 students who neither indicated that rating consistency nor comment helpfulness influenced the reviewing grade. Therefore, data from 195 students were analyzed using one-way, between-subjects ANOVAs comparing the three perceived conditions. Least significant difference (LSD) *post hoc* tests were used to determine which conditions were significantly different. To demonstrate the size of the difference, Cohen's *d* (i.e. mean difference divided by average standard deviation) for each comparison was included in the figures. Error bars in the figures represent one standard error.

Similar patterns were obtained when using assigned condition for analysis (see Table 2), but given how frequently students misperceived their assigned conditions, these effects were weaker. Because perceived conditions better represent the intended manipulation and the patterns were similar between assigned and perceived conditions, only the results for the perceived conditions will be discussed in detail. Initial analyses also included the condition in which students perceived that their grade was influenced by neither the rating consistency nor the comment helpfulness. In general, this condition was not significantly different from any of the three target conditions. Therefore, these analyses were not discussed.

### *Effects of accountability on ratings*

First, we examined the effect of accountability on the ratings students provided (see Figure 1). There were significant differences between the conditions, $F(2, 192) = 4.41$, $p = .01$. Students in the perceived both condition and students in the perceived comment only condition earned higher rating consistency scores than those in the perceived rating only condition, $t(123) = 3.0$, $p = .003$ and $t(96) = 2.3$, $p = .02$, respectively. Students in the perceived both condition did not earn significantly different rating consistency scores than those in the perceived comment only condition, $t(165) < 1$, $p = .40$. These results are more consistent with the depth-of-processing hypothesis.

**Table 2.** Assigned vs. perceived conditions: comparison of means.

| | Perceived condition | | | Assigned condition | | |
|---|---|---|---|---|---|---|
| | Both | Ratings | Comments | Both | Ratings | Comments |
| Volume of feedback | 1288 | 887 | 1336 | 1247 | 1072 | 1405 |
| Number of comments | 42.2 | 40.8 | 45.1 | 43.8 | 42.3 | 42.9 |
| Number of long comments | 6.8 | 2.0 | 7.0 | 5.8 | 4.3 | 8.3 |
| Number of criticism | 25.5 | 21.4 | 27.9 | 26.1 | 24.3 | 26.7 |
| Number of solutions | 22.4 | 17.9 | 24.4 | 22.8 | 20.7 | 23.7 |
| Number of localized comments | 16.6 | 12.7 | 18.2 | 17.6 | 13.6 | 18.2 |
| Rating consistency | 0.67 | 0.56 | 0.65 | 0.66 | 0.66 | 0.62 |

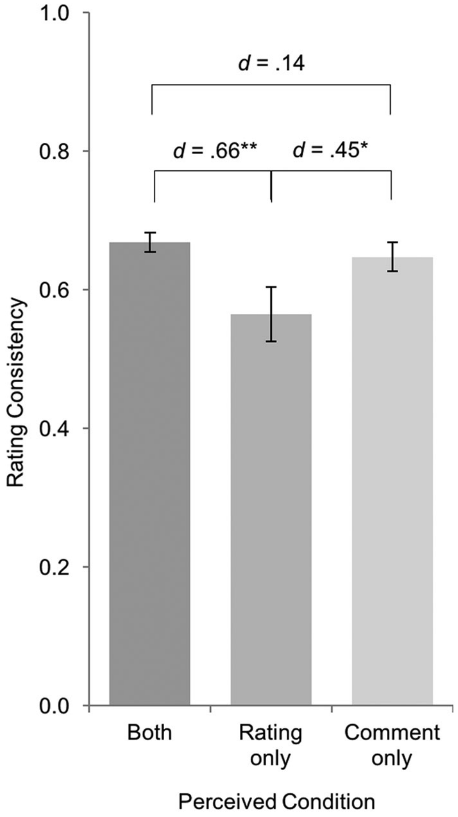

**Figure 1.** Rating consistency by perceived condition. **p* < .05. ** *p* < .01.

Given the differences in rating consistency between the conditions, one may wonder whether one condition was harsher than another (i.e. provided lower ratings on average). Such differences could produce confounds (e.g. on ICC via ceiling effects or on amount of comments via differential perceived amount of problems to comment upon). However, there were no differences in the average ratings provided between the conditions, *F*(2,192) = 1.61, *p* = .20.

### *Effects of accountability on comments*

Next, we examined the effect of accountability on the amount of comments students provided (see Figure 2). There were no differences in the number of comments provided, *F*(2, 192) = 1.06, *p* = .35, with most students providing only the minimal number of required comments. However, there were significant differences between the conditions for volume of feedback and number of long comments, *F*(2, 192) = 3.65, *p* = .03 and *F*(2, 192) = 3.94, *p* = .02, respectively. Students in the perceived both condition and students in the perceived comments only condition had higher mean volume of feedback than those in the perceived ratings only condition, *t*(123) = 2.4, *p* = .02 and *t*(96) = 2.6, *p* = .01, respectively. Similarly, students in the perceived both condition and students in the perceived comments only condition had more comments with 50 words or more than those in the perceived ratings only condition, *t*(123) = 2.6, *p* = .01 and *t*(96) = 2.6, *p* = .01, respectively.

Next, we analyzed the three automatically coded feedback features (see Figure 3). There were differences in the amount of criticism, solutions, and localized comments provided, *F*(2, 192) = 3.09, *p* = .05, *F*(2, 192) = 3.31, *p* = .04, and *F*(2, 192) = 2.36, *p* = .10, respectively. Students in the perceived comments only condition provided more criticism, more solutions, and marginally more

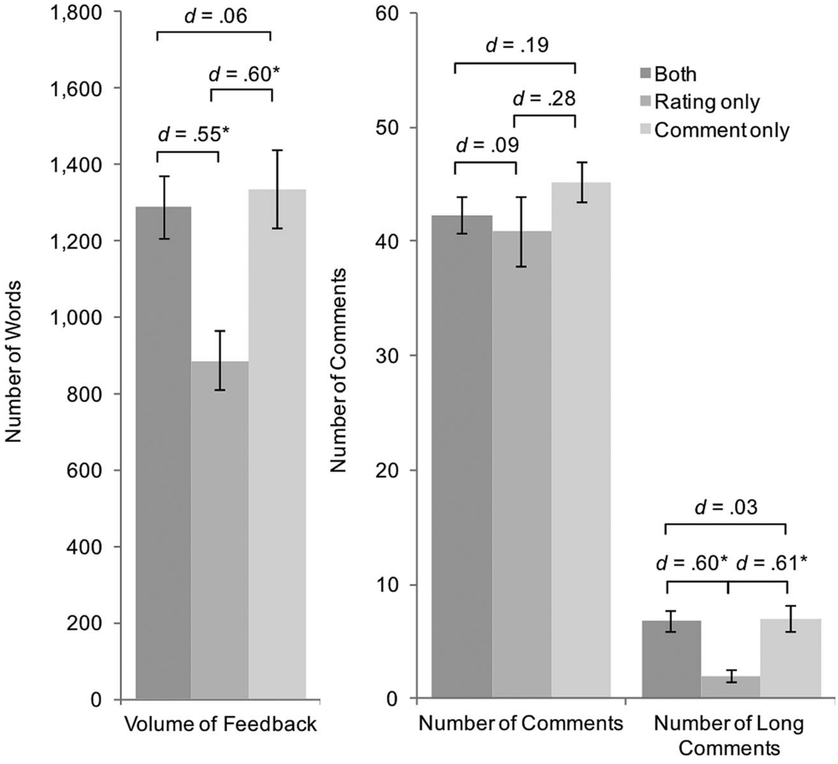

**Figure 2.** Amount of feedback (i.e. total volume of feedback, total number of comments, total number of long comments) by perceived condition. *p < .05.

localized comments, than those in the perceived ratings only condition, *t*(96) = 2.5, *p* = .01, *t*(96) = 2.6, *p* = .01, and *t*(96) = 2.2, *p* = .03, respectively.

Despite these differences, authors perceived the comments from all conditions to be equally helpful (*M* = 4.4, SD = 0.40), *F*(2, 192) = 1.19, *p* = .31, perhaps influenced by a ceiling effect in helpfulness ratings (i.e. most ratings were either 4 or 5 on the 5-point scale).

## General discussion

### *Summary of results*

The goal of the current study was to examine the effects of rating-focused versus comment-focused forms of accountability on the quality of peer assessment. We tested two hypotheses: (1) According to the *direct accountability hypothesis*, students who think their reviewing grade is influenced by the consistency of their ratings will more consistently rate their peers' work, while those who think their reviewing grade is influence by the helpfulness of the comments will provide more helpful feedback, and (2) According to the *depth-of-processing hypothesis*, students who think their reviewing grade is influenced by the helpfulness of their feedback will not only provide more helpful feedback but also more consistently rate their peers' work. Using different configurations for calculating the reviewing grade, students were explicitly held accountable for providing consistent ratings, constructing helpful feedback, or both.

The results from the current study did not support the direct accountability hypothesis – that is, students who thought their reviewing grade was influenced by the consistency of their ratings did not provide more consistent ratings than those who thought their reviewing grade was influenced

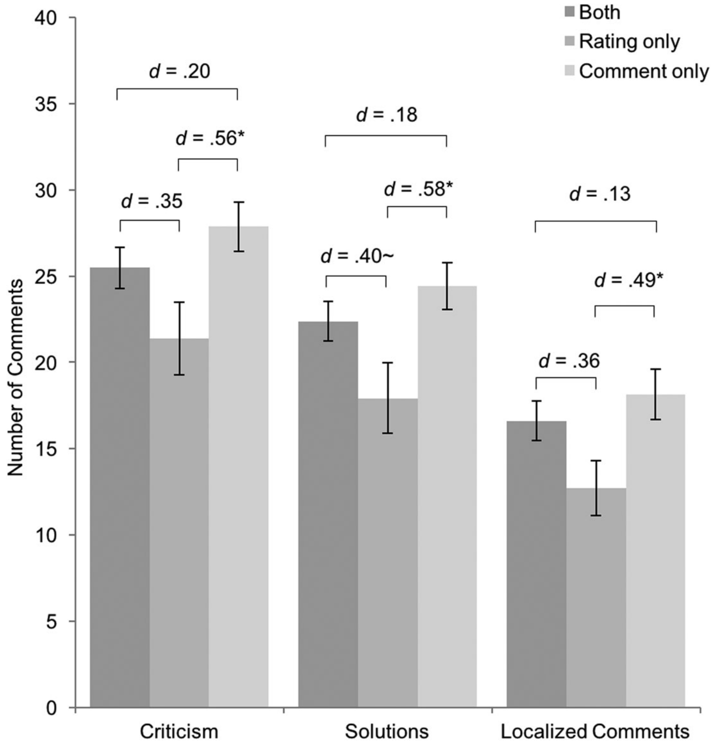

**Figure 3.** Feedback features (i.e. total number of criticism comments, total number of solution comments, total number of localized comments) by perceived condition. *$p < .05$; ~$p < .10$.

by the helpfulness of their comments. Rather, the results supported the depth-of-processing hypothesis. Feedback accountability improved the quality of the comments provided by reviewers (i.e. increased the volume and number of long comments, and sometimes the total number of criticism, solutions, and localized comments). Moreover, producing higher quality comments may have a stronger influence on the consistency of ratings than assigning a reviewing grade that reflects the rating consistency – that is, although providing comments may have an effect on rating quality, when reviewers are held accountable for producing higher quality comments, the effect is even more distinct. One possible mechanism that could explain these results relates to dual-process theory (Evans 2011), with greater attention to commenting producing a more deliberative, slow rating process.

### *Caveats and future directions*

The current study's findings are consistent with prior research that has demonstrated that constructing feedback is an important contributor to helping students learn how to write – rather than just evaluating the quality of a peer's work (Lu and Law 2012; Wooley et al. 2008). Future studies would be better positioned to directly examine theoretical explanations by collecting additional data, including surveys, additional performance measures, more carefully controlled peer review objects, and additional experimental contrasts (e.g. no accountability condition, alternative variations of accountability). More specifically, future work should directly examine potential cognitive mechanisms for the depth-of-processing hypothesis (e.g. whether there is a shift from fast, intuitive reasoning to slow, deliberate reasoning in completing ratings). More data are needed that directly speaks to whether reviewers process the rating task differently than the commenting task (e.g. intuitive feelings vs. a more analytic process).

One lesson learned from this study was that accountability was not easily manipulated through in-class instruction. In the currently study, only 30% of the students perceived the manipulation as intended. One possible explanation for this difficulty could be that the instructor cultivated a unique classroom culture by repeatedly emphasizing the purpose of peer assessment as a way to 'help peers improve the quality of their paper.' In doing so, most of the students (69%) believed their reviewing grade was based on the helpfulness of their comments. Therefore, it is important to see whether these findings replicate in contexts where the instructor does not make this emphasis. Moreover, students may differentially react to using grades as a motivator. Future research should incorporate a measure of students' perceived importance of grades that could be used as a covariate. In addition, one could measure whether the students believed the proportion of the final grade allocated to reviewing is a sufficient incentive (e.g. 3% in the current study). Another factor that could be related to students' motivation and rating consistency is their confidence of their rating and comments ability – that is, does that accuracy of students' judgments about their rating consistency and comment helpfulness depend on explicit accountability (e.g. deeper processing involved in commenting accountability)? Additionally, future research should also explore other ways to manipulate accountability that do not involve grades (e.g. leaderboards in PeerWise; Denny, Luxton-Reilly, and Hamer 2008).

The current study also exposed a few open questions. For example, future research should explore why students in the perceived both condition provided comments with only slightly more criticism, solutions, and localization than those in the perceived ratings only condition. In addition, future research could examine why authors did not perceive the differences in comment quality. It is possible that authors are not good at judging helpfulness, so future work should examine the validity and reliability of back reviews. Additionally, the current study only addresses the reliability of ratings. Future comparisons should also focus on the validity of ratings as addressed in CPR (Balfour 2013; Russell 2004). Finally, students may be concerned that if authors grade the reviews of their work, then the authors could retaliate for a critical review by giving low ratings to the reviewer. The SWoRD system accounts for this possibility by presenting the ratings in aggregate, but it is still possible for students to observe how much criticism is provided by a particular reviewer. In Mobius SLIP, the environment accounts for this possibility by requiring authors to rank reviewers rather than rate them (Gehringer 2014). A closer examination of how these methods affect the quality of the reviews would be useful.

## Conclusion

As Luxton-Reilly (2009) observed, less than half of the web-based, peer assessment systems included explicit accountability, and those that did varied in focus and approach. The findings from the current study support the use of feedback accountability via grades based on author's perception of how helpful they thought the comments they received were. This approach is consistent with multiple existing systems, including CrowdGrader, MobiusSLIP, Peer Grader, PECASSE, SWoRD, and Eli Review. Not only does this approach improve the quality of the peer feedback, it also increases the reliability of the peer ratings.

What is still unknown is how beneficial reliability grades alone will be. Again, multiple systems utilize this approach, including Aropä, CPR, Expertiza, and SWoRD. However, given the effect of feedback accountability on peer ratings, how this approach compares to a condition where neither the ratings nor the comments were held accountable is still unknown. Such a condition would not make practical sense in a classroom, and many teachers would likely resist using a peer assessment system that did not hold students accountable for their work.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Babik, D., E. F. Gehringer, J. Kidd, F. Pramudianto, and D. Tinapple. 2016, June. "Probing the Landscape: Toward a Systematic Taxonomy of Online Peer Assessment Systems in Education." Paper presented at the Workshop at Educational Data Mining (EDM), Raleigh, NC.

Balfour, S. P. 2013. "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review TM." *Research & Practice in Assessment* 8: 40–48.

Bruffee, K. A. 1980. *A Short Course in Writing: Practical Rhetoric for Composition Courses, Writing Workshops, and Tutor Training Programs*. Boston, MA: Little, Brown.

Cho, K., and C. MacArthur. 2011. "Learning by Reviewing." *Journal of Educational Psychology* 103 (1): 73–84. doi:10.1037/a0021950.

Cho, K., and C. D. Schunn. 2007. "Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System." *Computers & Education* 48 (3): 409–26. doi:10.1016/j.compedu.2005.02.004.

Cho, K., C. D. Schunn, and D. Charney. 2006. "Commenting on Writing: Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts." *Written Communication* 23 (3): 260–94. doi:10.1177/0741088306289261.

Cho, K., C. D. Schunn, and R. W. Wilson. 2006. "Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives." *Journal of Educational Psychology* 98 (4): 891–901. doi:10.1037/0022-0663.98.4.891.

de Alfaro, L., and M. Shavlovsky. 2014, March. "CrowdGrader: A Tool for Crowdsourcing the Evaluation of Homework Assignments." Paper presented at the 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA.

Denny, P., A. Luxton-Reilly, and J. Hamer. 2008. "The PeerWise System of Student Contributed Assessment Questions." Paper presented at the Tenth Conference of Australasian Computing Education (ACE 2008), Wollongong, Australia.

Elbow, P. 1973. *Writing Without Teachers*. New York: Oxford University Press.

Evans, J. S. B. T. 2011. "Dual-process Theories of Reasoning: Contemporary Issues and Developmental Applications." *Developmental Review* 31: 86–102. doi:10.1016/j.dr.2011.07.007.

Falchikov, N., and J. Goldfinch. 2000. "Student Peer Assessment in Higher Education: A Meta-analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (3): 287–322. doi:10.3102/00346543070003287.

Gehringer, E. F. 2014. "A Survey of Methods for Improving Review Quality." In *New Horizons in Web Based Learning: ICWL 2014 International Workshops, SPeL, PRASAE, IWMPL, OBIE, and KMEL, FET, Tallinn, Estonia, August 14-17, 2014, Revised Selected Papers*, edited by Y. Cao, T. Väljataga, J. K. T. Tang, H. Leung, & M. Laanpere, 92–97. Springer International.

Grabill, J., B. Hart-Davidson, and M. McLeod. 2012. "Whitepaper: Eli Review." http://elireview.com/wp-content/uploads/2014/12/elireview_whitepaper_final.pdf.

Hamer, J., K. T. K. Ma, and H. H. F. Kwong. 2005. "A Method of Automatic Grade Calibration in Peer Assessment." Paper presented at the Seventh Australasian Computer Science Education Conference (ACE'2005), Newcastle, Australia.

Hartberg, Y., A. B. Gunersel, N. J. Simspon, and V. Balester. 2008. "Development of Student Writing in Biochemistry Using Calibrated Peer Review." *Journal of the Scholarship of Teaching and Learning* 2 (1): 29–44.

Kaufman, J., and C. Schunn. 2011. "Students' Perceptions About Peer Assessment for Writing: Their Origin and Impact on Revision Work." *Instructional Science* 39 (3): 387–406. doi:10.1007/s11251-010-9133-6.

Lu, J., and N. Law. 2012. "Online Peer Assessment: Effects of Cognitive and Affective Feedback." *Instructional Science* 40 (2): 257–75.

Luxton-Reilly, A. 2009. "A Systematic Review of Tools that Support Peer Assessment." *Computer Science Education* 19 (4): 209–32.

Moffett, J. 1968. *Teaching the Universe of Discourse*. Boston, MA: Houghton Mifflin Company.

Nguyen, H. V., and D. Litman. 2014, June. "Improving Peer Feedback Predictions: The Sentence Level Is Right." Paper presented at the 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), ACL 2014 Workshop, Baltimore, MD.

Nguyen, H. V., W. Xiong, and D. Litman. 2014, June. "Classroom Evaluation of a Scaffolding Intervention for Improving Peer Review Localization." Paper presented at the 12th International Conference on Intelligent Tutoring Systems (ITS), Honolulu, HI.

Patchan, M. M., D. Charney, and C. D. Schunn. 2009. "A Validation Study of Students' End Comments: Comparing Comments by Students, a Writing Instructor, and a Content Instructor." *Journal of Writing Research* 1 (2): 124–52. doi:10.17239/jowr-2009.01.02.2.

Patchan, M. M., C. D. Schunn, and R. J. Clark. 2011. "Writing in Natural Sciences: Understanding the Effects of Different Types of Reviewers on the Writing Process." *Journal of Writing Research* 2 (3): 365–93. doi:10.17239/jowr-2011.02.03.4.

Piech, C., J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. 2013, July. "Tunes Models of Peer Assessment in MOOCs." Paper presented at the 6th International Conference on Educational Data Mining (EDM 2013), Memphis, TN.

Ramachandran, L. 2013. "Automated Assessment of Reviews." Doctoral Diss., North Carolina State University, Raleigh, NC.

Russell, A. A. 2004. "Calibrated Peer Review: A Writing and Critical-Thinking Instructional Tool." In *Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering and Mathematics (STEM) Education*, 67–71. Washington, DC: American Association for the Advancement of Science.

Schoenfeld, A. H. 2007. “What Is Mathematical Proficiency and How Can It Be Assessed?” *Assessing Mathematical Proficiency* 53: 59–74.

Schunn, C. D. 2016. “Writing to Learn and Learning to Write Through SWoRD.” In *Adaptive Educational Technologies for Literacy Instruction*, edited by S. A. Crossley and D. S. McNamara, 243–59. New York: Taylor & Francis, Routledge.

Shute, V. J. 2008. “Focus on Formative Feedback.” *Review of Educational Research* 78 (1): 153–89. doi:10.3102/0034654307313795.

Song, Y., Z. Hu, and E. F. Gehringer. 2015, October. “Pluggable Reputation Systems for Peer Review: A Web-service Approach.” Paper presented at the IEEE Frontiers in Education Conference (FIE), El Paso, TX.

Song, Y., Z. Hu, E. F. Gehringer, J. Morris, J. Kidd, and S. Ringleb. 2016, June. “Toward Better Training in Peer Assessment: Does Calibration Help?” Paper presented at the Computer-Supported Peer Review in Education (EDM workshop), Raleigh, NC.

Tinapple, D., L. Olson, and J. Sadauskas. 2013. “CritViz: Web-based Software Supporting Peer Critique in Large Creative Classrooms.” *Bulletin of the IEEE Technial Committee on Learning Technology* 15 (1): 29–35.

Topping, K. J. 2005. “Trends in Peer Learning.” *Educational Psychology* 25 (6): 631–45. doi:10.1080/01443410500345172.

Topping, K. J., R. Dehkinet, S. Blanch, M. Corcelles, and D. Duran. 2013. “Paradoxical Effects of Feedback in International Online Reciprocal Peer Tutoring.” *Computers & Education* 61: 225–31. doi:10.1016/j.compedu.2012.10.002.

Wooley, R. S., C. Was, C. D. Schunn, and D. Dalton. 2008. “The Effects of Feedback Elaboration on the Giver of Feedback.” Paper presented at the Cognitive Science, Washington DC.

Xiong, W., D. Litman, and C. Schunn. 2012. “Natural Language Processing Techniques for Researching and Improving Peer Feedback.” *Journal of Writing Research* 4 (2): 155–76.