When Trying to Be Helpful, Peer Reviews Are Also More Accurate

Melissa M. Patchan
West Virginia University
Allen Hall, room 504M
Morgantown, WV 26506
1-740-232-8670
melissa.patchan@gmail.com

Christian D. Schunn
University of Pittsburgh
3939 O'Hara Street, room 821
Pittsburgh, PA 15260
1-412-624-8807
schunn@pitt.edu

Russell J. Clark
University of Pittsburgh
3941 O'Hara Street, room 404
Pittsburgh, PA 15260
1-412-624-9204
ruc2@pitt.edu

ABSTRACT
To better understand why the positive effects of peer assessment are not consistent, we examined the influence of explicit accountability mechanisms in web-based peer assessment environments on the quality of peer ratings and peer feedback. More specifically, we tested two possible hypotheses (i.e., direct accountability hypothesis and depth-of-processing hypothesis) by comparing three accountability configurations: only rating accountability, only feedback accountability, or both rating and feedback accountability. From a large Introduction to Laboratory Physics course, 287 students' peer ratings and peer feedback were analyzed. Because student responses to a survey revealed that only 30% of the students perceived their assigned condition as intended, data was analyzed according to the perceived condition. Students who believed their reviewing grade was influenced by the helpfulness of their feedback were more likely to produce higher quality comments. More interestingly, these students also provided more accurate ratings than those who thought their reviewing grade was influenced by the accuracy of their ratings. These findings indicate that constructing helpful comments may have a stronger influence on student behavior than how a grade is assigned. Several suggestions for future directions were provided.

Keywords
peer assessment; peer ratings; peer feedback; accountability

1. INTRODUCTION
Peer assessment is the quantitative or qualitative evaluation of a learner's performance by another learner among students. It is typically implemented in classrooms with the intention of developing the knowledge or skill of all learners involved. Peer assessment has been supported by research for more than four decades [2, 7, 16]. This research has demonstrated that students are capable of providing valid ratings [8]. Moreover, students can provide feedback that is just as helpful as an instructor's feedback in helping their peers improve their drafts [20], and sometimes they can provide feedback that is more helpful [3, 4, 13]. However, as with most pedagogy, these effects are not consistent, which warrants a deeper investigation into the features of peer assessment. Therefore, the goal of the current study is to explore one of these features—that is, how explicit accountability mechanisms affect the quality of peer assessment.

Accountability for a given performance dimension occurs when a student is held responsible for the quality of completed work on that performance dimension. In a systematic review of tools that support peer assessment, Luxton-Reilly [15] observed that less than half of the web-based systems included some form of explicit accountability. These mechanisms varied in focus and approach. One accountability mechanism focused on overall motivation. For example, a leaderboard is used in PeerWise [6] to display who evaluated the most multiple-choice questions and the popularity of the peers' feedback. Other accountability mechanisms focus on the reliability or validity of the ratings. For example, an algorithm is used in Aropä [12] to calculate a reviewer weight, which indicates how similar the ratings a reviewer provided matches the ratings provided by other peers who evaluated the same work. However, Hamer and colleagues observed a wide range of reviewer weights that made it difficult to provide more specific suggestions to instructors about how to interpret the weights. To address the validity of ratings, reviewers complete a calibration task in Calibrated Peer Review [1, 18], which involves rating three essays strategically chosen by the instructor. Before reviewing the assigned peer documents, each reviewer must successfully complete this calibration task. In doing so, the accuracy of the reviewer's ratings is calibrated to the instructor. Success is dependent upon the effort put forth by the reviewer. Although a Reviewer Competency Index is calculated and used to weight the scores assigned to the authors, it is not used as a separate accountability mechanism (i.e., for grades). Finally, accountability mechanisms focus on the quality of the peer feedback. In Peer Grader [9] and PECASSE [11], authors evaluate the feedback that they received (also called metareviewing, back-review, double-loop feedback). This evaluation may involve rating the quality and helpfulness of the comments or whether the author agreed with the feedback. CrowdGrader [5] incorporates a grade based on the accuracy of the peer ratings as well as a grade based on the helpfulness of comments. In Expertiza [17], an automated metareview feature has been integrated, which automatically calculates 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. Also included was a metric for the number of unique comments provided. Despite the inclusion of these mechanisms, it is still unclear what effect the accountability mechanisms have on the

quality of peer assessment—that is, there has been no direct comparison between peer assessment with an explicit accountability mechanism versus without.

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) [4, 19]. This environment includes accountability mechanisms for both the peer ratings and peer feedback—that is, the grade students receive for completing the reviewing tasks comprises two parts: rating accuracy 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) with the assumption that these accountability mechanisms would directly affect the quality of peer assessment. This assumption led to our first hypothesis.

> *Hypothesis 1:* According to the **direct accountability hypothesis**, the accountability mechanisms directly affect the quality of peer assessment—that is, students who think their reviewing grade is influenced by the accuracy of their ratings are better positioned to consistently rate their peers' work than those who do not.

However, it is possible that only one of these accountability mechanisms is sufficient. Prior research on the benefits of providing feedback has demonstrated that reviewers who constructed feedback produced higher quality projects and essays than those who only rated the quality of the peers' work [12, 21]. These results indicate that constructing feedback involves deeper processing than just evaluating quality. In other words, reviewers only need to detect problems during the rating task, while reviewers need to detect, diagnose, and solve problems to construct helpful comments. These additional processes not only lead to more helpful comments, but after reflecting on their peers' errors, reviewers also perform better as authors on the task. Similarly, by better understanding the problems in a peer's text, the reviewer is expected to provide more accurate ratings. This assumption led to our second hypothesis.

> *Hypothesis 2:* According to the **depth-of-processing hypothesis**, only the feedback accountability mechanism is necessary—that is, students who think their reviewing grade is influenced by the helpfulness of their feedback are better positioned to consistently rate their peers' work than those who do not.

Note that both the direct accountability hypothesis and depth-of-processing hypotheses could be true—that is, students who think their reviewing grade is influenced by both the accuracy of their ratings and the helpfulness of their feedback are in a better position to consistently rate their peers' work, while students who think their reviewing grade is influenced by the accuracy of their ratings or the helpfulness of their feedback receive similar rating accuracy scores.

2. METHOD

2.1 Course Context & 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 recitation 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 recitations that were all taught by the same instructor and one 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 not perceive the manipulation as intended. Therefore, data from 287 students were included for analysis. This sample (52% female; M_{age} = 21.2 years; SD_{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).

2.2 Design & 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 newest version of the SWoRD system [19]. 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. 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 accuracy 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 accuracy 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 the accountability mechanism 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 five possible factors. Two items corresponded to the rating and commenting accountability mechanism (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). Three items were included as foils (i.e., whether my ratings are consistent with the ratings provided by the TA; the number of problems I find in my peers' papers; the length of my comments). Because 70% of students misperceived the accountability mechanism 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.

2.3 Measures

We examined several measures of rating accuracy and comment quality to better understand the effects of accountability on the peer review process.

2.3.1 Rating Accuracy

Each reviewer's *rating accuracy* 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 correlation between reviewer and peer means across the n = 40 ratings—4 papers x 10 rating dimensions).

2.3.2 Comment Helpfulness

The 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—that is, a mean of the received back-review ratings across all 24 comments they provided (4 papers x 6 comments / paper).

2.3.3 Amount of Feedback

The amount of feedback was examined using three measures that quantified the comments provided across the six reviewing dimensions for each of the four peers. First, the *length of feedback* was computed by summing the number of words across comments provided. 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 comments provided that consisted of 50 words or more (the threshold of 50 words was based on a frequency histogram, which revealed that the majority of the comments contained fewer than 50 words—i.e., comments that were 50 words or more were especially long).

2.3.4 Feedback Features

Three useful feedback features were automatically coded using a classification model derived from data mining and Natural Language Processing techniques [22]. 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.

3. Results & Discussion

3.1 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 accuracy 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 accuracy only. Interestingly, 50 (20%) students indicated that neither rating accuracy 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 the explicit accountability mechanisms on the quality of peer assessment, the main analyses focus on the perceived conditions, excluding the 49 students who neither indicated that ratings accuracy nor comment helpfulness influenced the reviewing grade. Therefore, data from 195 students was 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. Similar patterns were obtained when using assigned condition for analysis, but given how frequently students misperceived their

assigned conditions, these effects were weaker. Initial analyses also included the condition in which students perceived that their grade was influenced by neither the rating accuracy 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.

3.2 Effects on Accountability of 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 accuracy scores than those in the perceived rating only condition. Students in the perceived both condition did not earn significantly different rating accuracy scores than those in the perceived comment only condition. These results are more consistent with the depth-of-processing hypothesis.

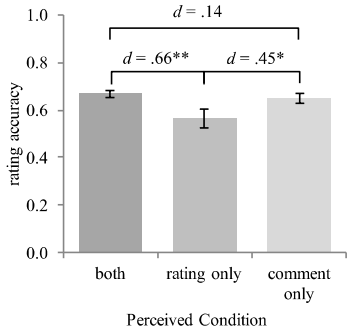

Figure 1. Rating accuracy by perceived condition.

*$p < .05$. **$p < .01$.

3.3 Effects on Accountability of Comments

Next, we examined the effect of accountability on the comments students provided (see Figures 2 and 3). Students in the perceived both condition and students in the perceived comments only condition provided longer comments than those in the perceived ratings only condition, $F(2, 192) = 3.65$, $p = .03$. Although there were no differences in the number of comments provided, $F(2, 192) = 1.06$, $p = .35$, students in the perceived both condition and students in the perceived comments only condition provided more comments with 50 words or more than those in the perceived ratings only condition, $F(2, 192) = 3.94$, $p = .02$. Next, we analyzed the three feedback features that were automatically coded [22]. Students in the perceived comments only condition provided more criticism, $F(2, 192) = 3.09$, $p = .05$, more solutions, $F(2, 192) = 3.31$, $p = .04$, and more localized comments, $F(2, 192) = 2.36$, $p = .10$, than those in the perceived ratings only condition.

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).

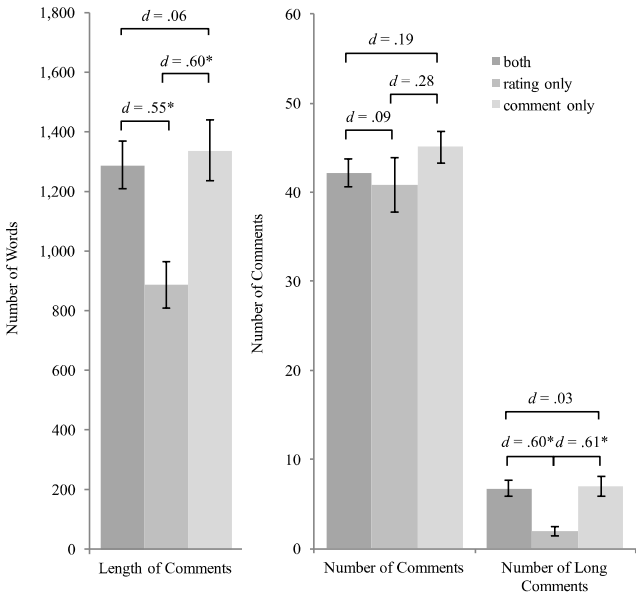

Figure 2. Amount of feedback (i.e., total length of comments, total number of comments, total number of long comments) by perceived condition.

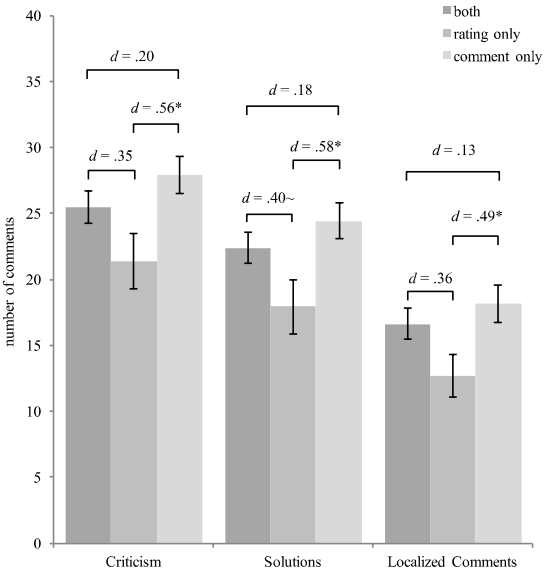

Figure 3. Feedback features (i.e., total number of criticism comments, total number of solution comments, total number of localized comments) by perceived condition.

*$p < .05$. ~$p < .10$.

4. General Discussion

The goal of the current study was to examine the effects of two explicit accountability mechanisms on the quality of peer assessment. Using different configurations for calculating the reviewing grade, students were explicitly held accountable for providing accurate ratings, constructing helpful feedback, or both. The results from the current study supported the depth-of-processing hypothesis. The feedback accountability mechanism improved the quality of the comments provided by reviewers (i.e., increased the length 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 accuracy of ratings than assigning a reviewing grade that reflects the rating accuracy—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. These findings are similar to those that demonstrate that constructing feedback is the largest source of learning to write (rather than just evaluating the quality of a peer's work) [14, 21]. 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).

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). Additionally, future research should also explore other ways to manipulate accountability that do not involve grades (e.g., leaderboards in PeerWise [6]).

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 Calibrated Peer Review [1, 18]. Finally, students may be concerned that if authors grade their 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 [10]. A closer examination of how these methods affect the quality of the reviews would be useful.

As Luxton-Reilly [15] observed, less than half of the web-based, peer assessment systems included explicit accountability mechanisms, and those that did varied in focus and approach. The findings from the current study support the use of a feedback accountability mechanism via grades based on author's perception of how helpful they thought the comments they received were.

5. REFERENCES

[1] Balfour, S. P. 2013. Assessing writing in MOOCs: Automated essay scoring and Calibrated Peer Review™, *Res. Pract. As.*, 8 (Summer 2013), 40-48.

[2] Bruffee, K. A. 1980. *A short course in writing. Practical rhetoric for composition courses, writing workshops, and tutor training programs*. Little, Brown and Company, Boston, MA.

[3] Cho, K., and MacArthur, C. 2011. Learning by reviewing, *J. Educ. Psychol.*, 103, 1 (Feb. 2011), 73-84. DOI=http://dx.doi.org/10.1037/a0021950.

[4] Cho, K., and Schunn, C. D. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system, *Comput. Educ.*, 48, 3 (Apr. 2007), 409-426. DOI=10.1016/j.compedu.2005.02.004.

[5] de Alfaro, L. and Shavlovsky, M. 2014. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. (Atlanta, GA, March 5-8, 2014). SIGCSE '14. ACM, New York, NY, 415-420. DOI=10.1145/2538862.2538900 http://doi.acm.org/10.1145/2538862.2538900

[6] Denny, P., Luxton-Reilly, A., and Hamer, J. 2008. The PeerWise system of student contributed assessment questions. In *Proceedings of the the 10th Conference of Australasian Computing Education* (Wollongong, Australia, January 22-25, 2008). ACE '08. ACS, Sydney, Australia, 69-74.

[7] Elbow, P. 1973. *Writing without teachers*. Oxford University Press, New York.

[8] Falchikov, N., and Goldfinch, J. 2000. Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks, *Rev. Educ. Res.*, 70, 3 (Fall 2000), 287-322. DOI=10.3102/00346543070003287.

[9] Gehringer, E. F. 2008. Strategies and mechanisms for electronic peer review. In *Proceedings of the 30th ASEE/IEEE Frontiers in Education Conference*, (Kansas City, MO, October 18-21, 2000). FIE '00. F1B 2-7. DOI=10.1109/FIE.2000.897675.

[10] Gehringer, E. F. 2014. A survey of methods for improving review quality. In *New Horizons in Web Based Learning*, Y. Cao, T. Väljataga, J. K. T. Tang, H. Leung, and M. Laanpere, Eds. Springer International Publishing, 92-97.

[11] Gouli, E., Gogoulou, A., and Grigoriadou, M. 2008. Supporting Self-, Peer-, and Collaborative-Assessment in E-Learning: The Case of PEer and Collaborative ASSessment Environment (PECASSE), *J. Interact. Learn. Res.*, 19, 4 (Oct. 2008), 615-647.

[12] Hamer, J., Ma, K. T. K., and Kwong, H. H. F. 2005. A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th Australasian Computer Science Education Conference* (Newcastle, Australia, Jan/Feb, 2005). ACE '05. ACS, Sydney, Australia, 67-72.

[13] Hartberg, Y., Gunersel, A. B., Simspon, N. J., and Balester, V. 2008. Development of Student Writing in Biochemistry Using Calibrated Peer Review, *J. Scholarship Teach. Learn.*, 2, 1 (Feb. 2008), 29-44.

[14] Lu, J., and Law, N. 2012. Online Peer Assessment: Effects of Cognitive and Affective Feedback, *Instr. Sci.*, 40, 2 (Jul. 2012), 257-275. DOI=10.1007/s11251-011-9177-2.

[15] Luxton-Reilly, A. 2009. A systematic review of tools that support peer assessment, *Comput. Sci. Ed.*, 19, 4 (Dec. 2009), 209-232. DOI=10.1080/08993400903384844.

[16] Moffett, J. 1968. *Teaching the universe of discourse*. Houghton Mifflin Company, Boston, MA.

[17] Ramachandran, L. 2013. *Automated assessment of reviews*. Doctoral Thesis. North Carolina State University.

[18] 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*. American Association for the Advancement of Science, Washington DC, 67-71.

[19] Schunn, C. D. in press. Writing to learn and learning to write through SWoRD. In *Adaptive Educational Technologies for Literacy Instruction*, S. A. Crossley and D. S. McNamara, Eds. Taylor & Francis, Routledge, New York, NY.

[20] Topping, K. J. 2005. Trends in Peer Learning, *Educ. Psychol.*, 25, 6 (Dec. 2005), 631-645. DOI=10.1080/01443410500345172.

[21] Wooley, R. S., Was, C., Schunn, C. D., and Dalton, D. 2008. The effects of feedback elaboration on the giver of feedback. In *Proceedings of the 30th Annual Conference of Cognitive Science* (Washington, DC, Dates, 2008). CogSci '08. Cognitive Science Society, Austin, TX, 2375-2380.

[22] Xiong, W., Litman, D., and Schunn, C. 2012. Natural language processing techniques for researching and improving peer feedback. *J. Writing Res.* 4, 2 (Oct. 2012), 155-176. DOI=http://dx.doi.org/10.17239/jowr-2012.04.02.3.